Philipp Berner, BSc

Technology Screening – Development of a Structured Approach for the Early Identification of Emerging Technologies and Startups

Master's Thesis

Angestrebter akad. Grad

Diplom-Ingenieur

Master's degree programme: Production Science and Management

submitted to

Graz University of Technology

Supervisor

Dipl.-Ing. BSc Matthias Helmut Friessnig

Dipl.-Ing BSc Hugo Daniel Karre

Institute of Innovation and Industrial Management
Univ.-Prof. Dipl.-Ing. Dr.techn. Christian Ramsauer

Graz, Februar 2017

Vorwort

Am Anfang steht immer eine Idee. Eine Idee als Ursprung von allem. Besonders wenn ich bedenke wie fragil eine anfängliche, oft noch nicht ausgesprochene, Idee ist; wie einfach sie zu einem Kompromiss verfälscht wird oder schlicht und einfach wieder untergeht; fasziniert mich der Gedanke, dass Ideen von Einzelpersonen erst durch den Austausch im Team zur Innovation werden.

So begann auch diese Arbeit als Idee und wurde zu etwas Konkreten. Ein Werkzeug, dass zukunftsträchtige Technologien & Startups aufspüren kann. Nach anfänglicher Planlosigkeit aufgrund der sehr vage definierten Aufgabenstellung wurde durch die erste Literaturrecherche schnell mein Interesse an dem Thema Technologie Management geweckt. Mit Fortschreiten der Arbeit nahm das Konzept immer mehr Form an und gab den ersten Blick auf das noch entfernte Ziel frei.

Einige Monate später; auch diese Masterarbeit, begonnen als Idee, bedurfte Teamwork zur Fertigstellung. Nicht nur während der anfänglichen Orientierungsschwierigkeiten, sondern während der gesamten Arbeit, konnte ich immer auf Hilfe meines Betreuer Matthias Friessnig zählen. Daher danke an dich, deine Ideen, deine Unterstützung und dein Engagement für dieses Projekt.

Ganz besonders bedanke ich mich bei meiner Familie, die mir mein Studium ermöglicht hat. Nicht nur der finanzielle, viel mehr der familiäre Rückhalt machten meine Studienzeit zu einem Lebensabschnitt, auf den ich gerne mit einem Lächeln auf den Lippen zurückblicken werde.

Auf dass noch viele Ideen zur Innovation werden.

Philipp

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master‘s thesis dissertation.

…… ……

Date Signature

Abstract

Due to the rapid technological development, especially since the mid-20th century, the amount of available information has increased exponentially. Before and at the beginning of this trend, the major problem was that information was not available or traceable. The continuing exponential data growth transformed this problem. Now most of the workload is shifted from data finding to filtering and verifying the data sources. Screening in general is the filtering and verifying extension of this aforementioned finding process. Hence, the emphasis of technology screening is set on raising the efficiency of finding useful and arising technologies, within all fields of application.

Rising labor expenses are the driving force for more efficiency on production lines. This development already starts to diffuse into the research and development (R&D) departments, with methods such as model-kits, R&D-cooperation's and so on. The technology screening takes this diffusion of higher efficiency one step further towards the start of every product development. Technology screening already in the earliest phases of product development enables discovering synergies, interest-clusters, helpful networks and many more easily. By means of fast and cost efficient methods, a rapid overview on emerging technology concepts, trends and startups is given. Thus, inefficient double developments as well as knowledge gaps in the field technology and market are identified and counteractions can be set.

In the venture capital market, technology screening is a state-of-the-art technology, despite being generally driven by one's network and professional relationships. In line with this master's thesis, a screening routine is designed, which will as a first step support the personal driven approach described above and, as a second step, replace it to enable more candor.

As input, only already available data are used, which forms one underlying difference to traditional methods. These data source searches are fully automated and the text semantics of the single documents are analyzed. Thus, a rapid overview on arising technology concepts, trends and startups is generated fast and cost-efficiently. In doing so knowledge gaps in the field of technology and market are identified and counteractions can be set.

The result of this automated data-processing routine is a report with a manageable amount of technology concepts, which are summarized in a technology catalog.

Kurzfassung

Durch die anhaltende technologische Entwicklung speziell seit Mitte des 20 Jahrhunderts ist die Menge der allgegenwärtig verfügbaren Informationen exponentiell gestiegen. In der Vergangenheit bestand das Problem, dass Informationen nicht erhältlich bzw. auffindbar waren; heute hingegen begegnet man einer Flut an Informationen zu den komplexesten Themengebieten, die online öffentlich verfügbar sind. Dadurch hat sich über die Zeit auch das Problem gewandelt: Anstatt die Information an sich zu suchen, müssen heute große Datenmengen durchsucht, gefiltert und auf Plausibilität geprüft werden, um zu den gewünschten Informationen zu gelangen.

Mit steigenden Lohnkosten stieg auch der Wunsch nach höherer Effizienz in der Produktion. Diese Entwicklung beginnt bereits durch Methoden wie Baukastensysteme, R&D-Kooperationen u. dgl. auch in die Produktentwicklung zu diffundieren. Das Technologie-Screening treibt diesen Diffusionsprozess im Streben nach Effizienz noch weiter an den Start der Produktentwicklung. Mit dem Technology-Screening können bereits in der frühesten Phase der Produktentwicklung Synergien, wie beispielsweise Interessenscluster oder wichtige Kontakte entdeckt werden. Es wird mit schnellen und kostengünstigen Methoden ein rascher Überblick über derzeitige Technologiekonzepte, Entwicklungen und Startups gegeben.

Technology Screening ist im Venture Capital Market eine bereits etablierte Methode, jedoch meist durch Netzwerke und Beziehungen getrieben. Im Rahmen dieser Masterarbeit wird eine Screening-Routine entworfen, die diesen oft von einzelnen Personen abhängenden Prozess zunächst unterstützen und im Idealfall ablösen und damit objektiver machen soll.

Als Input sollen lediglich bereits vorhandene Daten verwendet werden, wodurch sich dieser Ansatz grundlegend von traditionellen Methoden unterscheidet Diese Datenquellen werden vollautomatisiert durchsucht und die Semantik der einzelnen Textdokumente analysiert.

Somit wird mit schnellen und kostengünstigen Methoden ein Überblick über derzeitige Technologiekonzepte, Entwicklungen und Startups gegeben. Dadurch können zugleich ineffiziente Doppelentwicklungen sowie Wissenslücken im Bereich von Technologie und Markt frühzeitig erkannt und in Folge Gegenmaßnahmen gesetzt werden.

Als Ergebnis dieser automatisierten Datenverarbeitung entsteht Technologiekatalog.

Table of Content

1 Introduction

In this chapter, the initial situation, the objectives and the structure of this thesis are stated.

1.1 Initial Situation

Since the late-20th century, the access to data has been facilitated for the majority of people through the Internet. In former days, valuable data was just available for a selected group of people. Public libraries started changing the limitation of access to data and the Internet further fueled this development.

According to Weber et al., the Internet positively affects decision-making in politics, civil affairs and the industry. Additionally, the available semantic search algorithms easing the process of finding relevant data within huge unstructured datasets is a positive effect. In combination, this results in a far more efficient decision making process.[1]

The problem of data accessibility has transformed due to the rapid data growth of the Internet and its easy access to a new problem. The amount of data stored in the Internet is too big to be reviewed by traditional methods. Hence, the access to the data is no longer a limiting factor, but finding and filtering the right data has become the challenging part.

Today, more than ever companies and politics need to have the skills, tools and abilities to make use of fast-changing unstructured data. This skill is a key tool to obtain and develop technologies and innovate.[2]

Tools gaining benefit from the vast unstructured data amount of the Internet will facilitate many decision-making and development processes. The big challenge is to provide tools that are easy to use and fast in processing. Companies need tools that can be used by every employee in all different fields of application.

The available tools today look different. Typically forecasting processes rely on a scout network or experts as information source. The establishment of both a scout- and expert network requires resources, time and money. Additionally, these network systems are not designed and therefore not capable of taking benefit from the data

[1] Cf. Kang/Tsai/Horng (2009), p.1
[2] Cf. Amezcua-Martínez/Güemes-Castorena (2010), p.1

amount offered by the Internet. Resulting in several disadvantages of the network based approach[3]:

1. Expansive: A network of scouts is a big expense for companies with setup-, management- and maintenance cost on a monthly basis.
2. Limited effective range: A scout always has limited resources because of his network, being not equally distributed in all fields of application.
3. Biased: A scout as an individual is per se always biased on certain topics.

An alternative to this network approach is the expert approach. It is often performed as Delphi study, where interdisciplinary experts are surveyed several times with intermediate feedback loops about certain topics. By making use of the expert approach, similar drawbacks like with the network approach occur. Additionally, the expert approach is prone to be inert, because of the multiple questioning.[4]

This has led to the demand of a structured easy-to-use and fast tool, in a first instance to support traditional screening methods, and on a long-term basis substitute them. Hence, the developed approach should eliminate the above stated drawbacks by efficiently enlarging the effective range, not making use of expert interviews and finding unhyped technologies, showing a low probability to be found with conventional methods.

1.2 Research Objectives

The aim of this research is to show the feasibility of the developed screen routine approach. In addition, this thesis should form a good basement on which further research can be built on.

After studying relevant literature, important characteristics of the approach are stated. By observing these characteristics, a suited screen routine is developed and partly realized. The realization is limited by the framework of this thesis. As mentioned, the aim is to show the feasibility of the developed routine. In other words, the screen routine, if entirely programmed, would show the desired properties. Because of the limited resources of this research, some constraints are set, which can be seen in chapter 5.

As an output of this research, a technology catalog is generated, where several potential future trends are stated. All technologies within the technology catalog are

[3] Cf. Kochikar (2008), p.207
[4] Cf. Nagl (2016)

found by the routine. Therefore, the catalog constitutes a major part of the routine's feasibility.

In the end, research results will show an approach for the development of a screen routine. The routine needs to be designed according to the following needs:

1. Easy operation: Not only experts are able to use this routine.
2. Fast processing: The results of the routine need to be delivered in a sufficient time, in order to ensure the broad use of the routine.
3. Field of operation: The routine should be used in all different kinds of applications; therefore, it needs to be sufficiently flexible.

Especially the last point is important. The aim is to develop a tool suitable for a variety of applications and not only to be used in a niche. The following examples show some potential fields of application:

1. Venture capital market: To early identify new technology trends.
2. Competitor analysis: To get a fast overview of what competitors are developing.
3. Research and development (R&D): To early identify what is already on the market and showing the possibility of forming development clusters or cooperations.
4. Strategic planning: To identify new technology trends and substitutional technologies on time, to offer calls-for-action in terms of cooperate foresight.

1.3 Thesis Approach

The first sections of this thesis lay the theoretical foundation for the understanding and creation of the custom-developed screen routine.

Section 2 starts with the description of technology management in general. Afterwards the initial topic will be approached, the technology forecasting process. This process is shown in its single steps in the subsequent sections.

Since the amount of data needed to operate the screen routine is substantial, section 3 deals with the basic principles of Big Data. This is interesting for the screen routine, especially from a technical point of view, because sufficient hardware needs to be provided to fulfil the aimed requirements.

With section 4, the theoretical part ends. Because of the enormous data amount, which cannot be handled manually, data and especially text mining are key technologies for the success of the search routine's behavior. The basic principles and terms are explained in this section.

As mentioned, data sources are important. Section 5.3 shows the data source evaluation process executed within this research. The different criteria and thoughts are explained in detail. At the end, a combination of cluster criteria and a value benefit analysis were chosen to find 15 suited sources for the developed screen routine.

Section 5 shows the developed screen routine. Different prospected properties are listed and examples are shown. A holistic view of the process is given, to better understand the desired properties and results.

One output of the screen routine is the technology catalog, described and shown in section 5.5. In this catalog, technologies and trends are listed, all of which were found and processed by the routine.

Finally, a conclusion and an outlook on further development of the routine is given in section 6.

Figure 1.1 shows a simplified illustration of the developed screen routine approach. The orange marked steps are manually performed. The blue circle shows the automated search routine which, except the routine steering process, works automated. The output of the routine, the technology catalog, can be seen on the right. The number of performed search rounds can be adapted to the intended needs of the forecasting process. A more detailed description of this screen routine approach can be read in section 5.4.

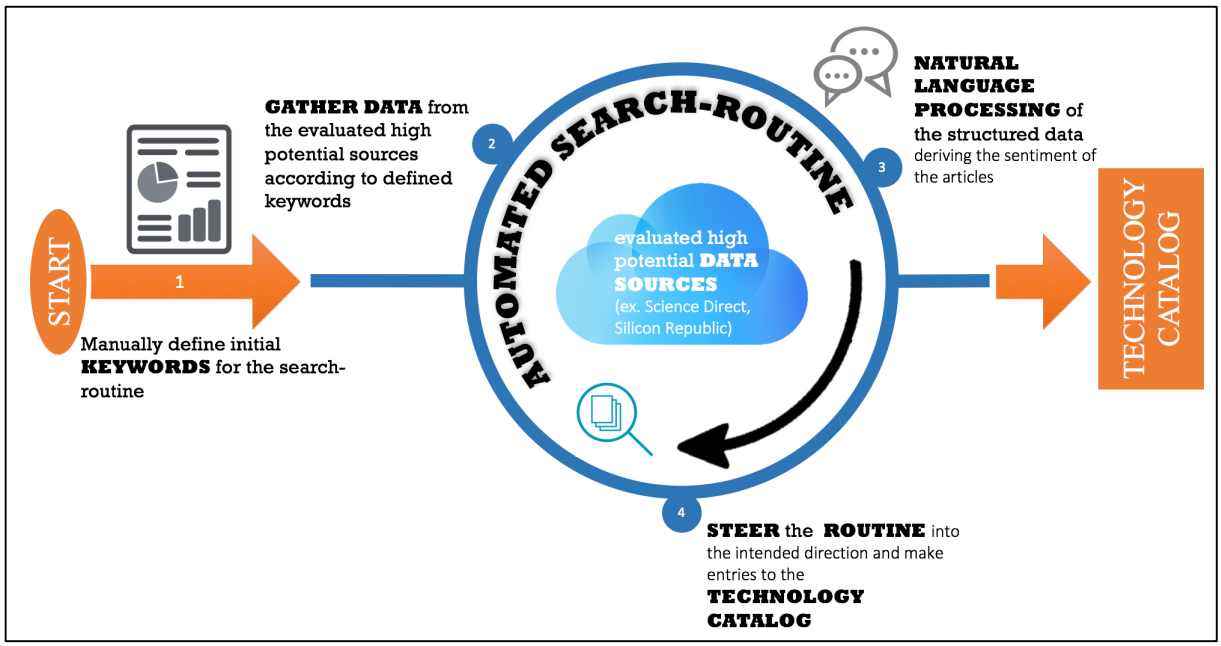


Figure 1.1: Simplified illustration of the screen routine approach

2 Forecasting as Part of Technology Management

Technology management is part of corporate management and acts on a medium- and long-term basis. The major task of technology management consists in all planning activities that ensure the long-term market success of a company. In addition to technological changes in products, changes regarding the production systems are also covered by technology management.[5]

As soon as emerging technologies replace a state of the art technology, those new technologies are called substitutional technologies. The early identification of substitutional technologies is one major part of technology management.[6]

For companies, substitutional technologies can either be a big opportunity or become a serious showstopper. The need for technology management greatly comes from these substitutional technologies. A good figure to represent this issue is the S-curve model, which will be described in chapter 2.1.

The very time for adopting a substitutional technology is crucial for the company's success. In practice, there is not one, but many potential substitutional technologies. Hence, scientists try to predict the performance of the different potential substitutional technologies by means of forecasting.[7]

A definition for technology forecasting is given by M. J. Cetron, who states that technology forecasting is:

'A prediction, with a level of confidence, of a technological achievement in a given time frame with a specified level of support'[8]*.*

Forecasting is nothing new, humans have always tried to predict the future. Only in the last few decades, this informal task has been given a structure. This analytical and structured way of forecasting is described by the above definition. The forecast has a confidence level, depending on the sources and the derived information. The time horizon is stated also and should be on a company's medium- to long-term basis.[9]

Generally, these forecasting activities are performed by closed circles of experts. Each of these experts predicts the future based on his/her single point of view, which results in a low level of confidence in the forecast. There are methods available trying to

---

[5] Cf. Schuh/Klappert (2011), p.5
[6] Cf. Kochikar (2008), p.207
[7] Cf. Schuh/Klappert (2011), p.315
[8] Cetron (1969)
[9] Cf. Burgelman/Christensen/Wheelwright (2005), p.62

synergistically integrate a broader view with multiple experts from interdisciplinary professions. One example of such methods is the Delphi Method, described in section 2.4.1, where interdisciplinary experts are surveyed several times with intermediate feedback loops about certain topics.[10]

Chapter 2.3 shows the forecasting process and the subsequent chapters deal with the four basic activities within this process according to Shuh & Klappert.

2.1 The S-Curve Model

If technological progress depended on random events only, where it is not possible to establish a relationship between the degree of technological progress and time (shown in Figure 2.1), any forecasting activity would be impossible. By means of historical data, a pattern can be seen between the rate of technological progress and time, hence it not only depends on random events. This pattern shows discontinuities at certain points in time and can be interpreted as an S-curve.[11]

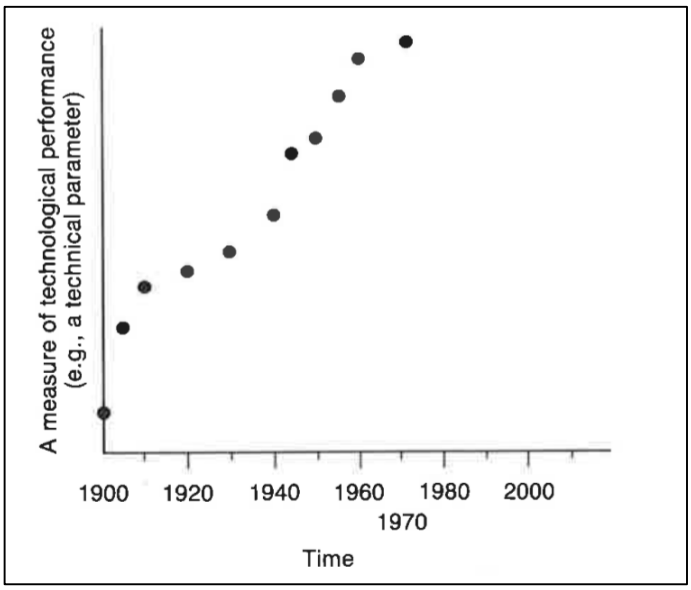


Figure 2.1: History of a developing technology, where the rate of advance does not follow any pattern[12]

Figure 2.2 shows the development process of a technology's performance as a function of the cumulative sum of the research and development (R&D) effort, which is equivalent to time. The curve marks several different types of technologies. At the very beginning of the R&D process, every technology starts as an embryonal technology. This stage is generally of less interest for companies, because of the high uncertainty

[10] Cf. Pietrobelli/Puppato (2015), p.2
[11] Cf. Burgelman/Christensen/Wheelwright (2005), p.67
[12] Burgelman/Christensen/Wheelwright (2005), p.67

about the technology developing as expected. Being further developed, a pacemaker technology is formed. The uncertainty of technological success is drastically lowered because of first industrial applications. This technology stage is located right at the edge of marketability. Getting more attention, a key technology is formed, by finding wide application within the branch. Key technologies often turn into a base technology quickly, which's performance potential is already being exploited. Technologies at this stage are very likely to be substituted by the next pacemaker technology.[13]

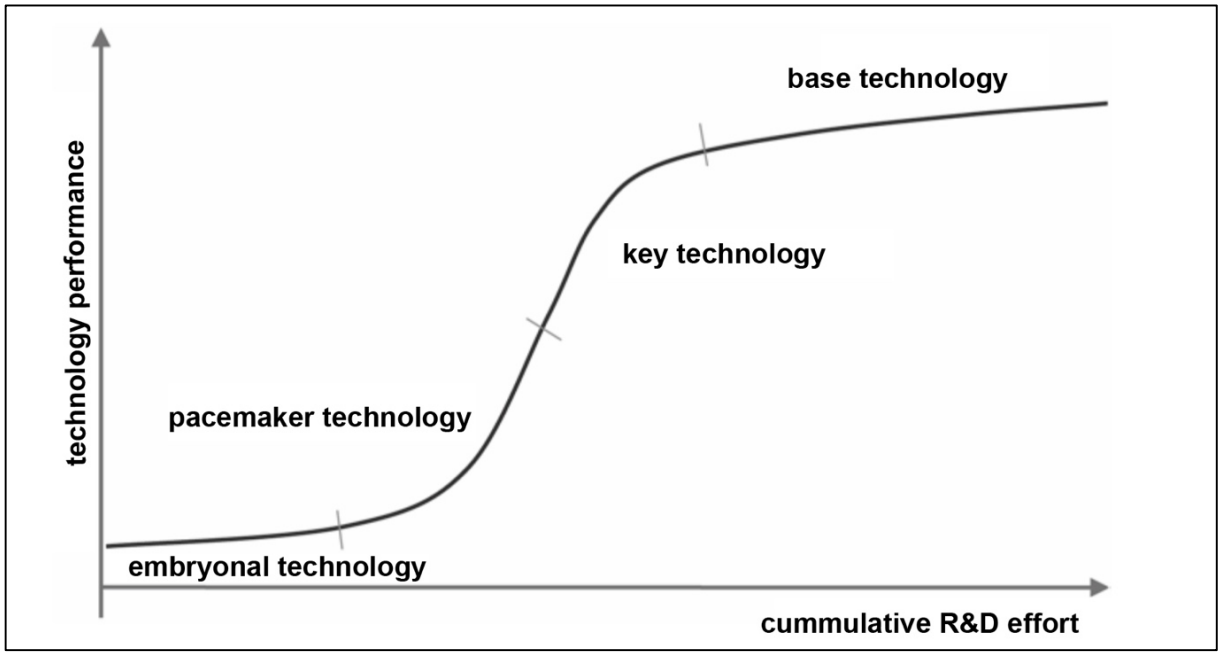


Figure 2.2: S-curve model with the different levels of maturity[14]

The S-curve model shows great similarity to a typical product life cycle, which can be seen in Figure 2.3. At the beginning, slow initial growth of performance (1) can be observed, which is followed by a rapid, exponential growth (2). Subsequently, the technology reaches physical limits and asymptotically approaches (3) the maximum performance. This asymptotical approach in performance is interesting for companies, because they mark the potential supremacy of a substitutional technology.[15]

Large companies tend to be inert in terms of disruptive technologies. Especially in times of disruptive change, such companies are in danger because of their slow decision-making processes and ignorance. This is the opportunity for small, agile new companies. A paradigm is formed. Large companies are way too slow to produce

[13] Cf. Schuh/Klappert (2011), p.43
[14] Schuh/Klappert (2011), p.43
[15] Cf. Burgelman/Christensen/Wheelwright (2005), p.67f

adequate responses in times of crisis and are therefore inferior to an agile small company.[16]

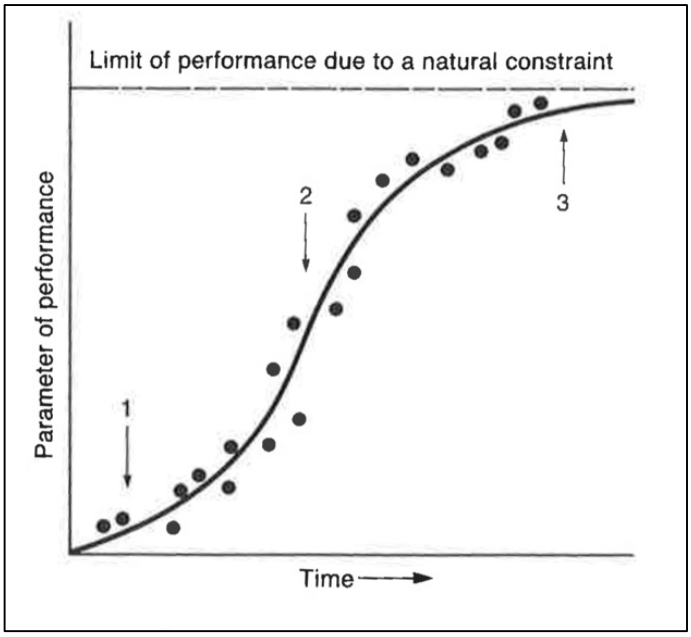


Figure 2.3: Different phases of the S-curve model[17]

2.2 Technology Forecasting Perspectives

For the technology forecasting process, two possible perspectives for the search of information can be observed:

1. Inside-out perspective (guided search)
2. Outside-in perspective (unguided search)

Both perspectives help to find different search fields and better understand the information demand determination as part of the forecasting process, described in the following chapter.[18] Both perspectives can be supported by different tools and methods. Researchers are steadily improving state of the art methods and developing new approaches for forecasting activities.[19]

Within the inside-out perspective, the search is kept in the company's environment and therefore deals with already established technologies. Because of this perspective, the

[16] Cf. Rohrbeck (2011), p.32
[17] Cf. Burgelman/Christensen/Wheelwright (2005), p.68
[18] Cf. Schuh/Klappert (2011), p.106
[19] Cf. Dobrzańska-Danikiewicz (2010), p.46

search fields are very specific and, regarding content, the information found will be near the key competences and value stream of the company, illustrated in Figure 2.4.[20]

For companies, the interpretation of weak signals plays a big role, since these signals are the indicators for change. This change can either be an opportunity or a risk for the company, depending on its preparation. The problem with the inside-out perspective is that these weak signals (likely developing out of a white spot) are often not spotted, because of concentrating on the internal problem statements. This is a benefit of the outside-in perspective.[21]

In contrast, the outside-in perspective searches not within the company's environment, that is, not in the already established or planned technology field. This perspective opens a very broad search, where the linkage of the technology's benefit to the company is established in the end, after having found a technology trend. Thus, the possibility for discovering white spots is generated, which can be seen in Figure 2.4. Scopes totally unknown to a company are called white spots. These white spots often develop into big opportunities, for example by giving a company the first mover advantage.[22]

Additionally, white spots and forecasting in general emphasize the need for change. Managers performing forecasting activities are forced to be aware of the constant changing environment, often resulting in an ease of actively adapting the company to the changing environment.[23]

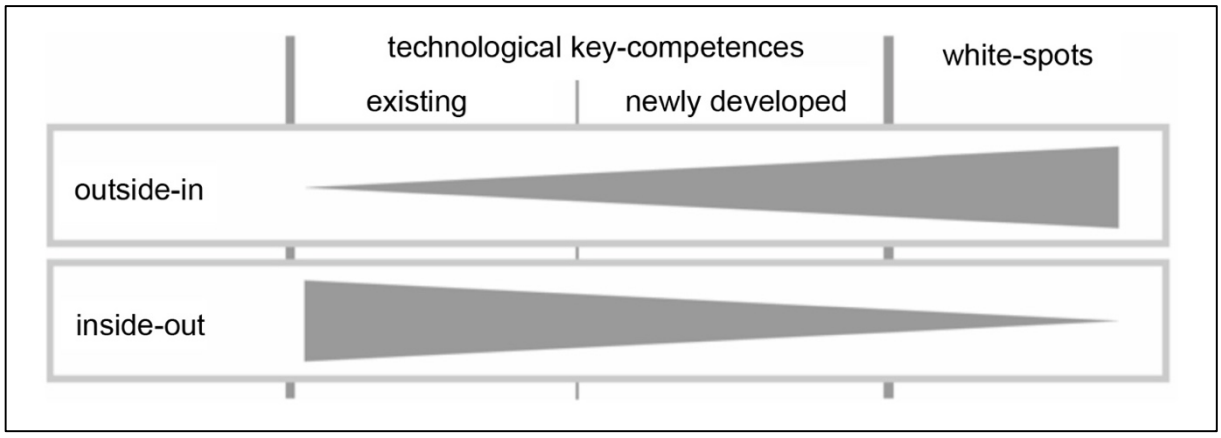


Figure 2.4: Outside-in and inside-out perspective of technology forecasting[24]

[20] Cf. Schuh/Klappert (2011), p.106
[21] Cf. Rohrbeck (2011), p.13f.
[22] Cf. Schuh/Klappert (2011), p.107
[23] Cf. Boe-Lillegraven/Monterde (2014), p.62
[24] Cf. Schuh/Klappert (2011), p.107

History shows that changes within firms are always characterized first by a long time of small incremental changes followed by a brief period of discontinuity and rapid change. The latter will happen anyway, it is part of the corporate foresight to ensure that the company is prepared for this rapid change and suitable technologies are ready to be applied.[25]

Nevertheless, in the forecasting process, both perspectives have the potential to discover new chances and risks and benefit from them. There is no universal strategy for finding the needed information, within or outside the company's environment. In practice, the inside-out perspective is the predominantly applied perspective. This results in the following major drawbacks of the outside-in perspective[26],[27]:

1. The effort to start an unguided search is far bigger than the effort of a guided one.
2. The unguided search has a lower probability to find applicable technologies that fit the company's needs.

2.3 Technology Forecasting Process

The technology forecasting process, considers the long-term future of science, technology, politics and society to form a profound base for decision making.[28] On a higher abstraction level, the aim of technology forecasting is to provide relevant information on changes inside and outside of the company's environment in time, to recognize potential threats and opportunities early enough to actively take measures to ensure the companies wellbeing.[29]

What is appropriate today in terms of skills and attributes, might be of less relevance tomorrow. The forecasting process ensures that a firm is prepared for this inevitability. More than that, it ensures that the change is appropriate and timely.[30]

For a company's strategic orientation, information on long-term technology trends, potential new business segments and substitutional technologies are of big interest. On that basis, internal key competences, key technologies and search fields to which relevant information should be gathered can be defined. The determination of the

[25] Cf. Rohrbeck (2011), p.16
[26] Cf. Schuh/Klappert (2011), p.107f.
[27] Cf. Schuh/Klappert (2011), p.108
[28] Cf. Pietrobelli/Puppato (2015), p.1
[29] Cf. Schuh/Klappert (2011), p.89
[30] Cf. Burgelman/Christensen/Wheelwright (2005), p.58

information demands should be done or at least organized by the persons in charge of the total technology forecasting process.[31]

For this reason, a structured and systematic approach describing how to gather, handle and analyze relevant information is needed. Schuh and Klappert describe the technology forecasting process, like depicted in Figure 2.5, as a continuous process consisting of the following four major activities[32].[33]

1. Information demand determination
2. Information sourcing
3. Information assessment
4. Communication of information

This forecasting process not only describes how forecasting activities are carried out, it also explains how and with which tools to implement it into the companies' framework.[34]

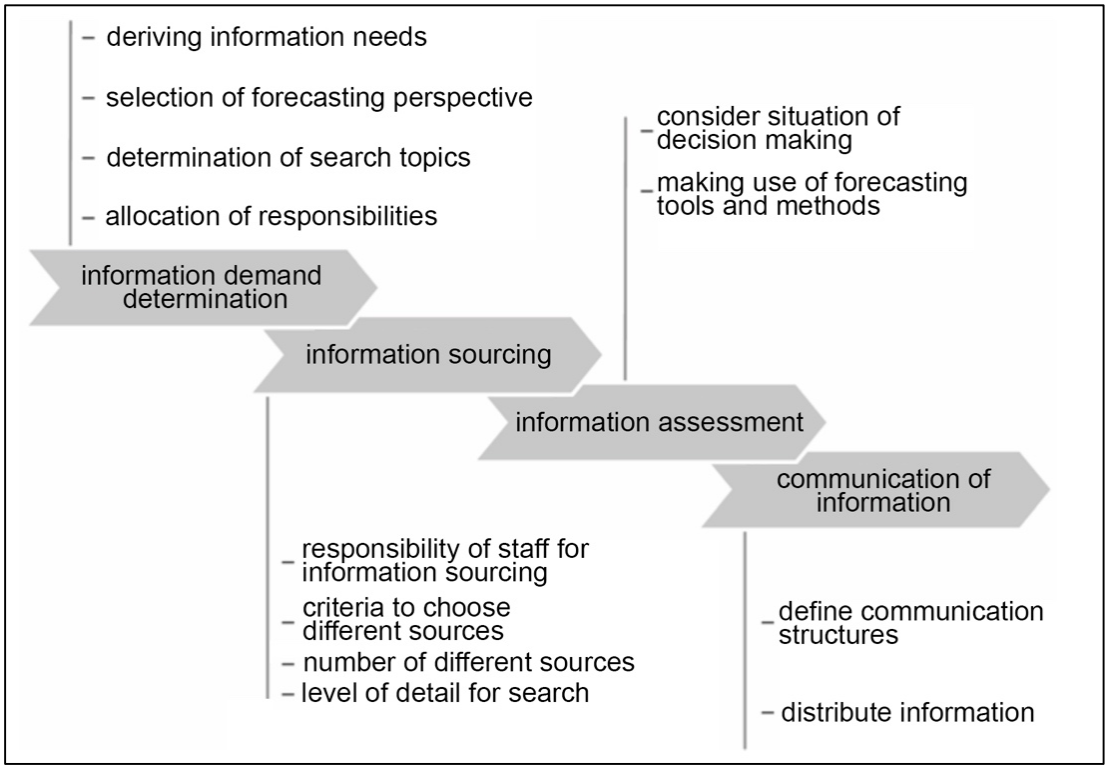


Figure 2.5: Technology forecasting process[35]

[31] Cf. Schuh/Klappert (2011), p.104
[32] Cf. Schuh/Klappert (2011), p.102f.
[33] Cf. Schuh/Klappert (2011), p.103
[34] Cf. Boe-Lillegraven/Monterde (2014), p.62
[35] Cf. Schuh/Klappert (2011), p.103

2.3.1 Information Demand Determination

The general orientation of the technology forecasting process is defined by the information demand determination, which is to circumscribe the guided as well as the unguided search for relevant information.[36]

According to Rohrbeck, the 5 Ws (who, what, when, where, why) are a good point to start. It helps to keep orientation while defining which information is needed.[37]

As a first step, the information demand needs to be deduced and be as detailed as possible, to later meet the requirements. According to the inside-out perspective, actual problems or missing key competences define the information demand. After having defined the information needs, the search perspective needs to be chosen, in other words, the orientation of the technology forecasting needs to be set. [38]

If the inside-out perspective is used, different search topics need to be chosen. This step, to a lesser degree, is also interesting for the outside in process, to later facilitate the subsequent allocation of responsibilities. The determination of search topics is essential when technology forecasting is performed with traditional methods. Because the available amount of information, especially because of the Internet, is too big to be gathered, handled and analyzed by traditional methods.[39]

Big data tools can help to overcome this limitation. By not simply string searching, but recognizing entities and concepts of articles, and websites, it becomes possible to process unmet big amounts of data. Although in this chapter the focus is on traditional forecasting methods, more information on Big Data tools can be found in chapter 3.4.[40]

The information demand determination phase ends by combining and merging the different information needs and search fields. Thematically similar information demands should be combined to ensure an efficient collaboration throughout the whole technology forecasting process.[41]

A good tool to communicate and visualize the different search fields is the monitoring radar.[42] There, the different search fields and the clear borders are visualized, as shown in Figure 2.6. Concentric circles depict the technology maturity. The nearer a

[36] Cf. Schuh/Klappert (2011), p.106
[37] Cf. Rohrbeck (2011), p.134
[38] Cf. Schuh/Klappert (2011), p.104
[39] Cf. Schuh/Klappert (2011), p.103f.
[40] Cf. Fasel/Meier (2016), p.269
[41] Cf. Schuh/Klappert (2011), p.118
[42] Cf. Rohrbeck (2006), p.1

technology is to the center, the more mature it is. In analogy to technology maturity, time itself can be chosen.[43]

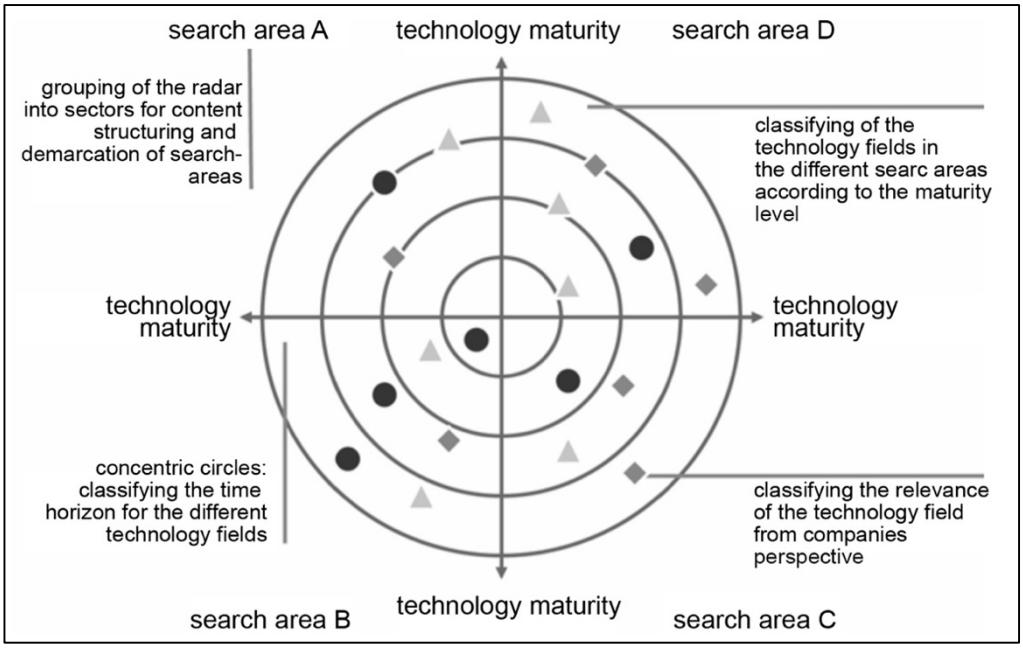


Figure 2.6: Elements of a monitoring radar[44]

2.3.2 Information Sourcing

Information sourcing is the binding element between information demand determination and information assessment. The goal of information sourcing is to have all the information needed and relevant ready for the subsequent processes. In order to structure and define the information sourcing, several parameters need to be determined[45]:[46]

1. Responsibility of work staff for information sourcing
2. Criteria for choosing different sources
3. Number of different sources
4. Level of detail for the search

The responsibility of who should perform the information sourcing depends on the aim of the forecasting process. If the forecasting should be processed on a detailed problem statement within the company, the sourcing needs to be performed by an expert in that field. By contrast, if the forecasting process is used to identify the next

[43] Cf. Schuh/Klappert (2011), p.118
[44] Cf. Schuh/Klappert (2011), p.119
[45] Cf. Schuh/Klappert (2011), p.121f.
[46] Cf. Schuh/Klappert (2011), p.121

mega trends or other influences of strategic interest, the sourcing needs to be carried out by a person operating at strategic levels.[47]

There are three possible inputs for the forecasting process. This is crucial due to the dependency of the results' quality on the input quality. Unfortunately, no future data is available, resulting in these three input types[48]:[49]

1. Information from the past
2. Knowledge of the present
3. Logical thought processes, insights and judgements

Choosing the right information sources is the key factor for a successful forecast. First of all, sources fitting the given topic need to be identified and access needs to be established. Defining source criteria is helpful for ranking the high number of possible sources. Those criteria can vary according to the given topic and intention. For example, research-oriented detail forecasting is best carried out drawing on sources such as articles and texts dealing with fundamental and applied science. By contrast, patents, legal norms and topic-unspecific articles are better suited to cover a broader context. As shown in Figure 2.7 emerging knowledge is always implicit, unpublished and local.[50]

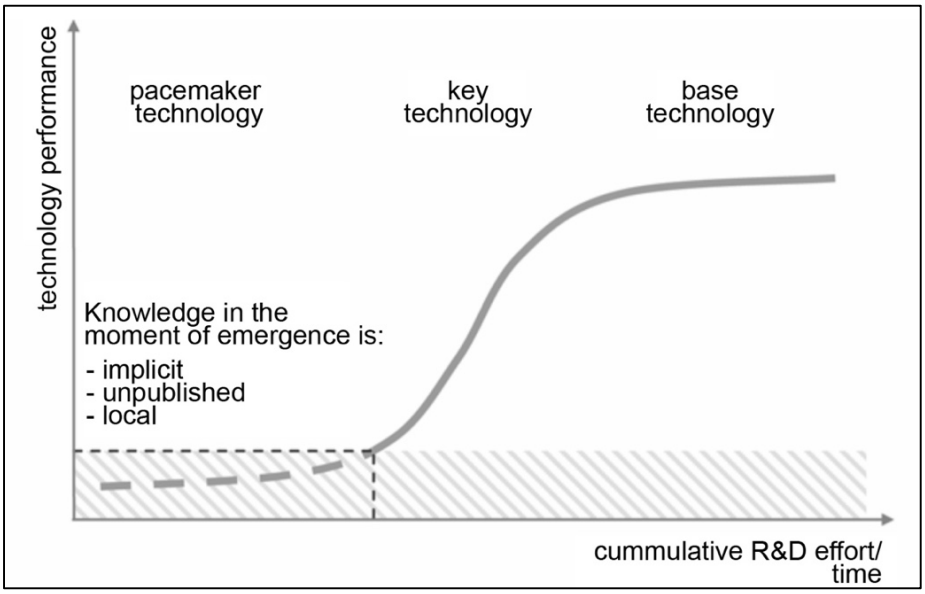


Figure 2.7: Characteristics of emerging knowledge[51]

[47] Cf. Schuh/Klappert (2011), p.121
[48] Cf. Burgelman/Christensen/Wheelwright (2005), p.65f.
[49] Cf. Burgelman/Christensen/Wheelwright (2005), p.65
[50] Cf. Schuh/Klappert (2011), p.122
[51] Cf. Schuh/Klappert (2011), p.123

Applying traditional methods, like scout networks or expert interviews, information sourcing requires high personnel costs. In addition, costs arise for information searching, fostering and updating of information sources and for services like subscriptions. Companies show efforts to lower these costs by using big data tools.[52]

The first step into lowering the costs of traditional methods is by introducing the e-foresight. This concept emerged in relation to already commonly known and used concepts like e-banking, e-commerce, e-management and many more. The concept of e-foresight is not directly part of the forecasting process but is a transition to the usage of Big Data tools within forecasting. E-foresight makes use of traditional forecasting tools (e.g. expert interviews, Delphi surveys, scenario planning) and implements IT-tools. This interdisciplinary merge makes it possible to online fill in surveys, make interviews and set appointments. Resulting in a higher cost effectiveness, because of lowering the traveling and time expenses.[53]

Figure 2.8 shows the proposed approach for information sourcing. First, the information demand needs to be clearly defined. After understanding the information demand, suitable sources are selected. Then, an appropriate approach for gathering the information is chosen. This approach needs to address the different sources and fulfil the overall forecasting goals. The last step is to gather the information applying the predefined parameters.[54]

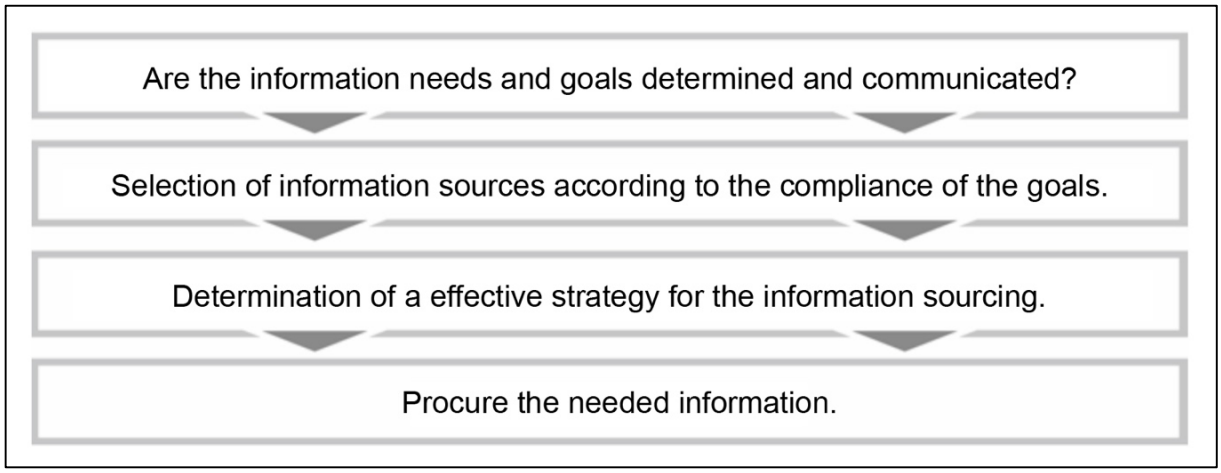


Figure 2.8: The information sourcing approach[55]

Several examples for different information sources are depicted in Figure 2.9.

[52] Cf. Schuh/Klappert (2011), p.124
[53] Cf. Dobrzañska-Danikiewicz (2010), p.42
[54] Cf. Schuh/Klappert (2011), p.129
[55] Cf. Schuh/Klappert (2011), p.129

Source		Environmental area			
Name	Description	Competitor	Customer	Political	Technological
Analyst reports	For trends analysis and financial data	✓	✓	✓	✓
Blogs	Online journals written by individuals	✓	✓	✓	✓
External experts	External people with specific knowledge	✓	✓	✓	✓
Internal experts	Employees with specific knowledge	✓	✓	✓	✓
Internet	Searched with standard search engines	✓	✓	✓	✓
Journalists	Specific group of experts who have domain knowledge and are well-connected	✓	✓	✓	✓
Newspapers and magazines	Particularly industry-specific magazines	✓	✓	✓	✓
Personal contacts	To gather informal information	✓	✓	✓	✓
Regional representatives	Product, marketing, and sales managers in regions	✓	✓	✓	✓
Statistical databases	OECD, World Bank, International Monetary Fund and from governments	✓	✓	✓	✓
Benchmark talks	Talks with direct competitors on non-competitive issues			✓	✓
Conferences and fairs	Trade fairs and for example technology conferences	✓			✓
Patents, publications	Accessed by databases and analyzed with specific software	✓			✓
Research reports	From public research projects such as EU- or nationally funded projects		✓		✓
Risk capital market	Tracking start-ups and private equity companies	✓	✓		✓
Scouts	Dedicated internal or external people hired to gather and disseminate information		✓		✓
Supplier and customer talks	Contacts to companies directly linked in the value chain		✓	✓	✓

Figure 2.9: Examples of information sources and the environmental area they are used in[56]

[56] Rohrbeck (2011), p.99

2.3.3 Information Assessment

The amount of information gathered during the information sourcing process is, despite information demand determination, substantial and thus further preprocessing and concentrating needs to be done. The assessment covers analyzing tools as well as prediction tools. Again, the approach chosen to process the information depends on the expected results of the forecasting process. The choice of the assessment approach and therefore the assessment methods depends on the goals, time horizon and the information basis of the forecasting process. Information assessment consists of the following three steps[57]:[58]

1. Selection: Reduction of the overall information amount by assessing relevance and priority.
2. Analysis: Condensing and analysis of information according to purpose and actual meaning.
3. Prediction: Interpretation of information in order to derive future developments and their meanings.

Figure 2.10 shows selected methods of information assessment. They can be categorized into qualitative and quantitative methods (plotted on the axis of ordinates). On the axis of abscissae, the time span for which the foresight process should be applied can be seen.

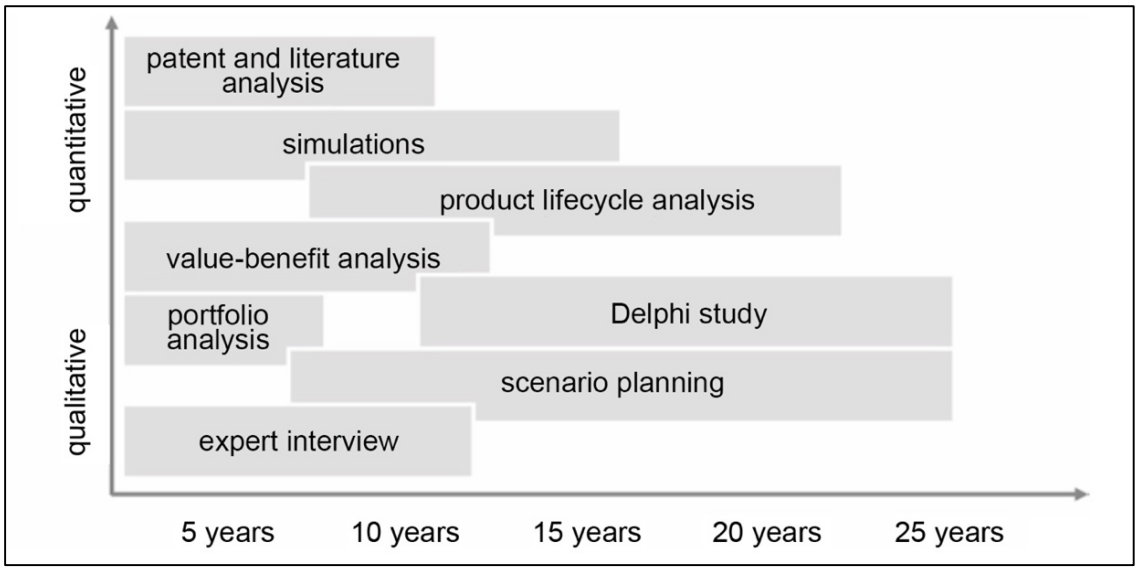


Figure 2.10: Selected methods used in technology forecasting[59]

[57] Cf. Schuh/Klappert (2011), p.130f.
[58] Cf. Schuh/Klappert (2011), p.131
[59] Cf. Schuh/Klappert (2011), p.134

The majority of companies that use foreasting methods do so by means of no specially developed IT tools. Especially quantitative methods (patent and publication analysis, simulations, …) are supported by IT tools, but as soon as reaching the qualitative methods, IT support is no longer used. This is because, for qualitative methods, tools are either too complex to use or no IT tools are available because of the flexibility offered by qualitative methods.[60]

Apart from medium- and long-term prediction and assessment of technological changes within the environment, technology assessment serves to analyze and prioritize concrete technologies and their alternatives. From the outcome, suggestions for action can be derived and communicated. This is the basis for many strategic decisions made in companies.[61]

Figure 2.11 shows the four elements that need to be included in every forecasting method.

Element	New	Description	Authors, year: page	Original name of element
Reach	✓	Describes how deeply a company scans; current business, adjacent business, and white spaces	Reger (2001b:539)	Introduces white spaces
Scope		Describes how broadly a company scans (technology, socio-cultural, customer, competitors, and political environment)	Becker (2002:15) Jain (1984:120)	Thematic areas Scope of scanning
Time horizon		Describes the time horizons of foresight activities (ranging from the near future to 30 years into the future)	Becker (2002:14–15)	Time horizon
Sources		Describes the sources of information; differentiated into internal vs. external, formal vs. informal	Jain (1984:124) Becker (2002:15)	Information sources Selection of sources of information

Figure 2.11: Key elements being part of every technology forecasting activity[62]

2.3.4 Communication of Information

For the efficient communication of information, it is not only necessary to know who it is handed to, but it is essential to establish a clear communication structure. In this structure, the way how information is communicated to the single persons is determined. Both ways, bottom-up and top-down, are used for communication. The latter, for instance, is used when people in charge of the forecasting process get to

[60] Cf. Möhrle (2008), p.50
[61] Cf. Schuh/Klappert (2011), p.135
[62] Rohrbeck (2011), p.75

know the search field or the corporate strategy. Nevertheless, bottom-up is the major direction of information flow during forecasting, when the results of the forecasting process are passed on to management and strategic level employees. The communication structure defines how to behave in routine situations and also in case of unscheduled events. The user knows exactly when to address whom in which kind of situation. This structure is essential for taking action on the spotted opportunities and risks.[63]

Apart from official communication, which is specified by the communication structure, room for unofficial communication between the employees needs to be established as well. This unofficial communication helps to define a common understanding of certain information, by talking on the different technologies and trends. Therefore, the processes of information assessment and information communication are strongly associated.[64]

2.4 Technology Forecasting Methods

As shown in Figure 2.10, there are different forecasting methods available and are chosen by the forecasting company depending on the aimed results. This chapter shows some best-practices forecasting methods, according to Schuh & Klappert (2011), that are used in companies.

2.4.1 The Delphi Method

Expert opinions are playing and will always play a big role in forecasting processes. An expert's opinion offers a short time lag and a high-quality source. Additionally, s/he can interpret weak signals and logically combine different scenarios and impacts. By means of expert interviews, very specific appraisals can be generated, but with the major drawback that results always reflect the opinion of a single individual.[65]

To overcome the major drawback of expert appraisals, the Delphi Method has been developed. This method eliminates the major drawback by using a questionnaire handed out to a panel of experts of diverse fields of specialty. A very important point is that the experts are not aware of the identity of their fellow members. The Delphi process consists of the following steps[66]:[67]

1. Round: The questionnaire is sent to the panel and returned by them via post.

---

[63] Cf. Schuh/Klappert (2011), p.136
[64] Cf. Schuh/Klappert, 2011, p.137
[65] Cf. Tidd/Bessant (2013), p.73
[66] Cf. Burgelman/Christensen/Wheelwright (2005), p.73f.
[67] Cf. Burgelman/Christensen/Wheelwright (2005), p.74

2. Round: The questionnaires of round 1 are analyzed and sent back to the expert panel. The experts are informed about the analysis and shown the average group results. Then, the experts are asked to correct their answer if they see fit. Experts showing 'extreme' answers in round 1, drifting far away from the average, are asked to reason their answers.
3. Round: The answers are analyzed again and the reasoning of the 'extreme' answers are stated and handed back to the panel. Then, the experts are asked to once again reconsider their replies.
4. Further rounds for clarification can be carried out if necessary.

2.4.2 Technology Roadmapping

Technology roadmapping can be seen as an universal tool for companies to reach their intended goals. Technology roadmapping ensures that the goals set are the right ones, and also defines the calls for action to reach these. A technology roadmap depicts development directions and its results in a timely manner. Furthermore, it shows technology fields, how technologies are developing according to different key factors, how and when technologies are building on each other or substituting each other, and so on.[68]

Technology roadmapping often gets mixed up with scenario planning, as both entail forecasting. But scenario planning focuses on the end result, whereas technology roadmapping also depicts the way to this result. For technology roadmapping, the result is broken down into single steps that need to be taken in a certain order. There are different technology roadmaps available[69]:[70]

1. Retrospective roadmap: This roadmap shows the development of a technology from the past until the present.
2. Prospective roadmap: This roadmap shows the development from the present situation into the future. For prospective roadmaps, a distinction in the topic's broadness can also be made. An explorative roadmap shows the development of a whole technology field, whereas, for example, a product roadmap only shows the development of the product's technology.

[68] Cf. Möhrle (2008), p.164
[69] Cf. Möhrle (2008), p.166f.
[70] Cf. Möhrle (2008), p.167

2.4.3 Scenario Planning

Scenario planning is based on a wide-range environmental analysis to define possible future scenarios. Generally, not only one scenario is analyzed, but a few of them. For scenario planning, other methods and tools from technology forecasting are used. Scenario planning is typically used by politics but is today also a state of the art forecasting method.[71]

Usually, scenario planning starts with finding the exogenous parameters influencing the topic under examination. Subsequently, assumptions are made and justified based on those parameters. Alternative parameters are assumed and then combined in a logical way. Through this, different scenarios are generated. A detailed step-by-step approach can be seen in Figure 2.12.[72]

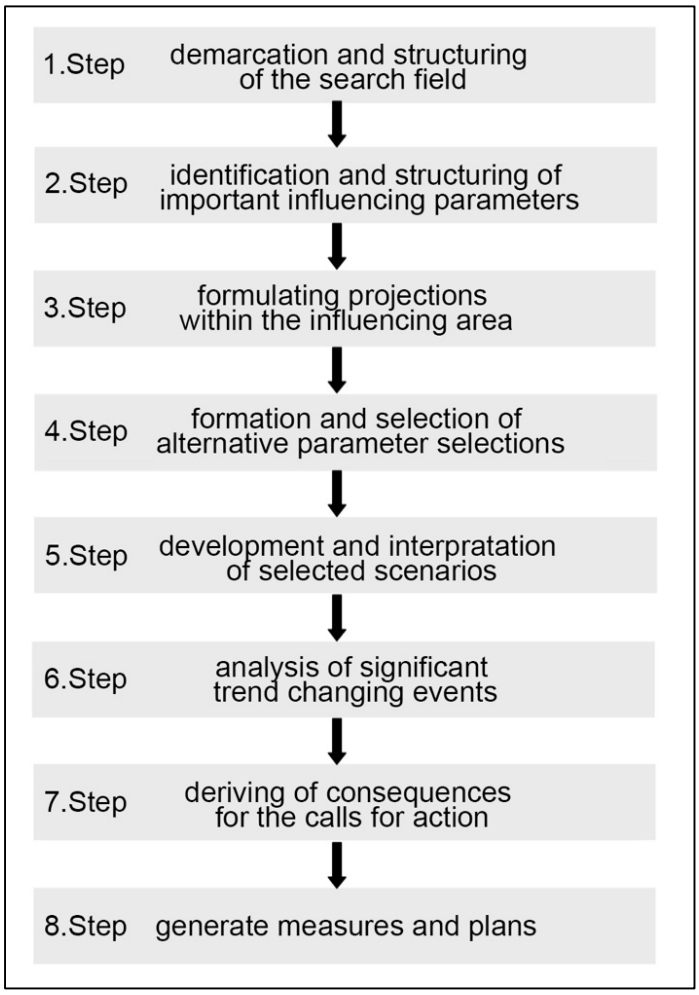


Figure 2.12: Eight step approach for scenario planning[73]

[71] Cf. Tidd/Bessant (2013), p.74f.
[72] Cf. Möhrle (2008), p.172f.
[73] Cf. Möhrle (2008), p.173

2.5 Technology Scouts as Part of Technology Management

In this chapter, the basic principles and ideas behind technology scouts are explained. Because scouts are the traditional method, the developed screen routine tries to support and, on a long-term basis, substitute the basic ideas behind technology scouts that are worth being shared.

Nevertheless, there are multiple search strategies for innovation, Tidd & Bessant give another overview of different methods, which can be seen in Table 2.1.[74]

Search Strategy	***Mode of operation***
Sending out scouts	Dispatch idea hunters to track down new innovation triggers.
Exploring multiple futures	Use futures techniques to explore alternative possible futures; and develop innovation options from that.
Using the web	Harness the power of the web, through online communities, and virtual worlds, for example, to detect new trends.
Working with active users	Team up with product and service users to see the ways in which they change and develop existing offerings.
Deep diving	Study what people actually do, rather than what they say they do.
Probe and learn	Use prototyping as mechanism to explore emergent phenomena and act as boundary object to bring key stakeholders into the innovation process.
Mobilize the mainstream	Bring mainstream actors into the product and service development process.
Corporate venturing	Create and deploy venture units.
Corporate entrepreneurship	Stimulate and nurture the entrepreneurial talent inside the organization.
Use brokers and bridges	Cast the ideas net far and wide and connect with other industries.
Deliberate diversity	Create diverse teams and a diverse workforce.
Idea generators	Use creativity tools.

Table 2.1:Search strategies to innovate[75]

[74] Cf. Tidd/Bessant (2013), p.78
[75] Cf. Tidd/Bessant (2013), p.278

2.5.1 Scout Definition and Motivation

Scouts are employees gathering information for the innovation process of a firm. This information gathering is usually done externally, which means that scouts are sent out and screen not only the company's environment, but globally for idea triggers that are relevant for the company's field of operation. They could look for technological triggers, emerging markets, or trends and competitor behavior. What they all have in common is the task to detect forms of innovation, often in the most unexpected places.[76]

Especially in the telecommunication industry, scouts are an established form of corporate foresight This is because the telecommunication industry has shown massive and multiple disruptive technological changes. Additionally, this industry is a rapidly changing business, where information does not last long. Many big companies failed to adapt to these massive disruptive technological changes and opened the way for smaller and agile companies taking the opportunity.[77]

These experiences in the telecommunication industry have created a deep-rooted awareness of the need of corporate foresight. The new agile companies became aware of their opportunities and formed a corporate culture where individual initiative was strongly empowered and formed the base for future scouts, increasing the chance of identifying new substitutional trends early and being able to produce effective calls for action in a timely manner.[78]

Having a very short time lag is the biggest advantage of technology scouts. Time lag is defined as the time spanning from the initial scientific discovery to the identification of a technology. This time lag can be of up to 18 to 24 months in publication and patent analysis. This time advantage is paid for with high costs for the establishment, management and maintenance of a scout network. Another drawback is the lack of scalability. The capacity of a single scout is limited, hence the only possibility to increase the output is to hire more scouts, which in turn increases fix costs in the management of the network.[79]

Figure 2.13 shows the typical setup of a scout network. A technology scout serves as an information node or an information hub in such networks. The task of the scout is to search for information within his/her own network and communicate the findings to

[76] Cf. Tidd/Bessant (2013), p.278
[77] Cf. Rohrbeck (2011), p.123
[78] Cf. Rohrbeck (2011), p.125
[79] Cf. Rohrbeck (2006), p.979

his/her direct contacts. There is also a tight network among the scouts. This has two reasons[80]:[81]

1. Validation of weak signals: The scouts need to communicate with other scouts to validate weak signals and decrease the danger of misinterpretation.
2. Information to trade: In order to stay an interesting contact for their sources, they need to be able to trade gathered information from other scouts.

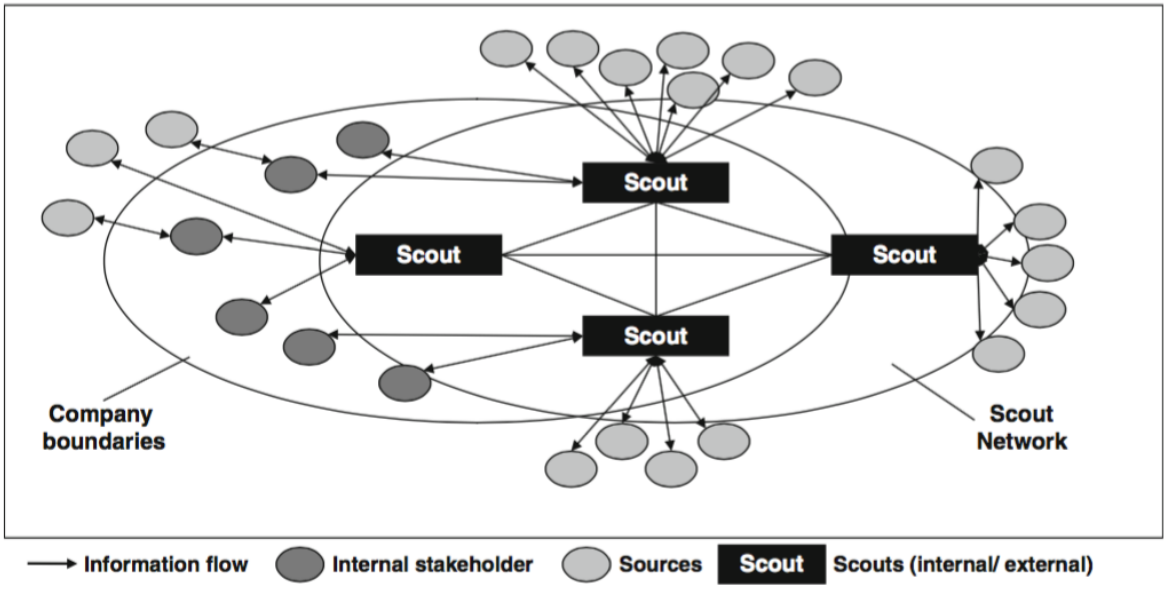


Figure 2.13: Structure of a scout network[82]

For technology scouting on a global scale, there is no other possibility than using external scouts. Typically, companies use both internal and external scouts, because external scouts offer limited knowledge of the needs for the internal stakeholders. This limits the usefulness of their information. There are three things to take into consideration when working with external scouts[83]:[84]

1. Scouts need to be familiar with the organization they work for, so they know how new issues are dealt with.
2. Scouts need to know how new information is channeled in the companies. There needs to be a clear structure establishing what information is reported to whom.
3. Scouts need to be informed about the innovation priorities of the company.

[80] Cf. Rohrbeck (2011), p.126f.
[81] Cf. Rohrbeck (2011), p.127
[82] Rohrbeck (2011), p. 127
[83] Cf. Tidd/Bessant (2013), p.278f.
[84] Cf. Rohrbeck (2011), p.127

Motivation plays a major role in scouting networks, since scouts are dependent on the favor of their sources to share the information with them. Hence, in a holistic view of the system, a win-win-win setup needs to be in place, with companies, scouts and sources as the winners. For this reason, each participant needs to have a motivation for his/her doing. The motivation for the company is detecting new technology trends and becoming aware of disruptive changes. For the sources and scouts, motivation is not as easy to provide, but an overview on the different motivations can be seen in Table 2.2.[85]

Actor in scouting networks	***Used incentives***
Internal scouts	Recognition Strengthening of internal network Monetary reward in bonus scheme
External scouts	Payment per relevant technology Payment of periodical fee Business development opportunity for consultants
Academic sources	Chance for joint research projects Recognition
Industry sources	Business development and sales Collaboration opportunity Validation of internal foresight insights

Table 2.2: Motivation sources in scout networks[86]

2.5.2 Scouting Rings

Some companies took the idea of scout networks one step further and formed scouting rings. In these scouting rings, scouts from different companies and fields of operation meet and exchange information. These meetings are institutionalized through virtual

[85] Cf. Rohrbeck (2011), p.128
[86] Rohrbeck (2011), p.129

and presence meetings. The rings are formed locally, with intersections to other companies. A typical setup of such a scouting ring can be seen in Figure 2.14.[87]

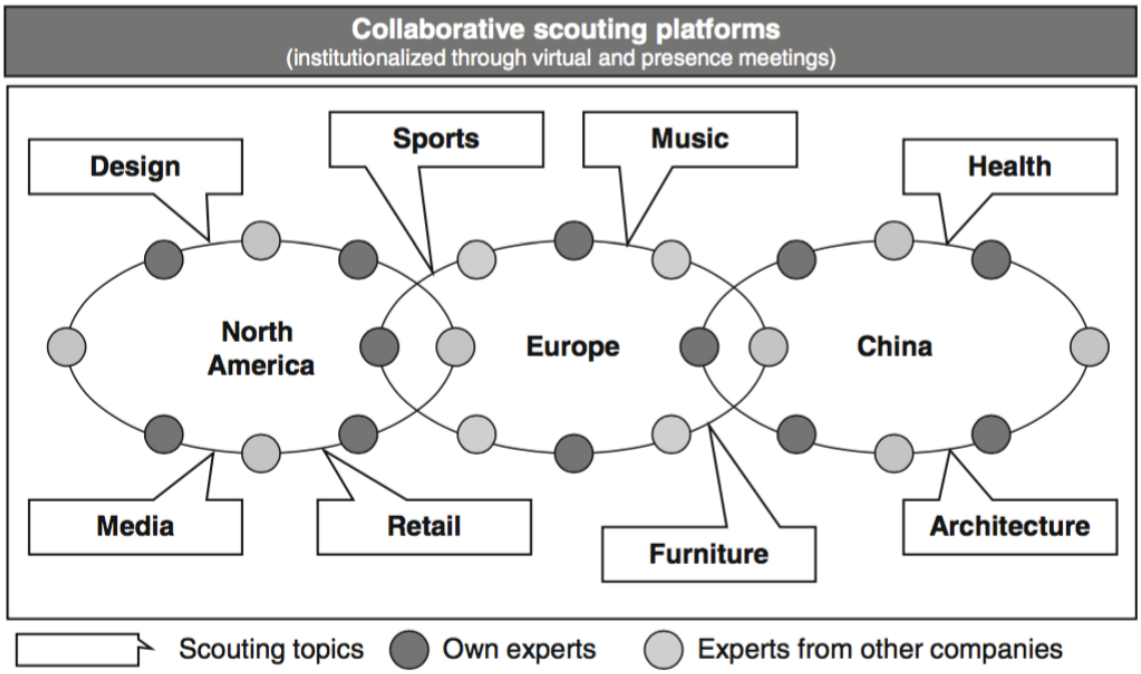


Figure 2.14: Example setup of a scouting ring[88]

Three primary drivers why such collaborative scouting activities with different companies emerged can be identified:[89]

1. It is expected that certain trends have an impact on all industries. For example, the trend toward health and sustainability-oriented lifestyles affects many different industries. Therefore, companies are interested in establishing a foresight mechanism to identify such issues.
2. All participating companies have the advantage of shared costs.
3. The rate of misinterpretations of weak signals can be reduced by multiple points of view. Especially when strongly consumer-driven companies can be brought together into a ring.

[87] Cf. Rohrbeck (2011), p.130f.
[88] Rohrbeck (2011), p.130
[89] Cf. Rohrbeck (2011), p.129

3 Big Data in Technology Management

Big Data is a much-discussed topic these days, probably because the global amount of data grows every second by more than 30.000 gigabytes. The magnitude of data is gradually exceeding common data handling systems.[90]

After giving a definition for Big Data, the Primary Data Circuit describing how this amount of data can be generated every second is explained. At the end of this chapter, technical and mathematical models that help deal with these huge data amounts are explained.

3.1 Definition of Big Data

Big Data can be defined as data that, because of its volume, is not storable, not analyzable and not computable with traditional hardware. Big Data analytics need data warehouses that provide real time access to unconventional amounts of data. This gets realized with NoSQL architectures and InMemory technology, described in the subsequent chapter 3.3. Moreover, the data characteristics are different. Big Data is much more heterogeneous than conventional data. This means that external data is also contemplable, as well as structured, semi- and unstructured data. In terms of Big Data analysis tools are no longer limited to structured data sets, also textual, as well as image or audio data can be processed. In addition, closed internal datasets in companies are opened to external data and broaden the company's view to a more global scale.[91]

Implementing Big Data tools offers a variety of benefits to companies. Some of these benefits are as follows:[92]

1. Information transparency is increased.
2. Higher frequency of data processing and analyzing.
3. Data shows a much higher level of detail and enables running more detailed simulations with more detailed results.

Big data is characterized by the 'big Vs'. According to different sources, the number of 'Vs' ranges from three to five. For the sake of consistency, all five characteristic 'Vs' of Big Data are stated in the following:[93]

---

[90] Cf. Marz/Warren (2015), p.1
[91] Cf. Fasel/Meier (2016), p.5
[92] Cf. Fasel/Meier (2016), p.6
[93] Cf. Iafrate (2015), p.3ff.

1. Volume: The data volume is extensive and lies in the range of Tera- and Zettabytes (1 Megabyte = 10^6 Bytes, 1 Gigabyte = 10^9 Bytes, 1 Terabyte = 10^{12} Bytes, 1 Petabyte = 10^{15} Bytes, 1 Exabyte = 10^{18} Bytes, 1 Zettabyte = 10^{21} Bytes)
2. Variety: Variety refers to the storage of structured, semi-structured and unstructured data. This means that multimedia data, like images, audio and video files also form part of these datasets.
3. Velocity: Velocity means that the data can be processed and analyzed in real time.
4. Value: Big Data applications are aimed at generating value for a company. Hence the applications are applied in working fields where a sufficientlever can be applied.
5. Veracity: Due to the opening of the internal datasets and taking advantage of external data sources, the quality of external data cannot always be ensured. This means that special algorithms need to be applied to assess and define the quality of different sources.

Many webbased companies, like Facebook[94], Amazon[95] and Google[96], process Petabytes' of data every day. Hence, these companies are dependent on Big Data tools to handle these data amounts. Companies not directly operating in the Information Technology (IT) branche, often do not see the benefit or value of Big Data tools for their purposes. This is due to the superficial knowledge about Big Data and the assumption that it only means dealing with great amounts of data. However, as shown, this is only a small portion of the whole concept.These traditional companies are often surprised by how much business-relevant information they are already gathering and what value can be generated from this data.[97]

3.2 Big Data Primary Circuit

The Big Data Primary Circuit can be described as data generator. The following list gives examples that illustrate which sources already feed Big Data:[98]

[94] https://www.facebook.com/
[95] https://www.amazon.com/
[96] https://www.google.com/
[97] Cf. Fasel/Meier (2016), p.4
[98] Cf. Bachmann/Kemper/Gerzer (2014), p.21f.

- Internet:
    - Unencrypted emails
    - Text messages
    - All kinds of postings on social media like Facebook[99], Twitter[100], …
    - Search engine queries
    - Order, buy and payment processes
    - All sorts of downloads
- Mobile devices
    - Data generated via apps
    - Smartphone location data
    - Communication and contact information
    - Images taken with these devices
- Digital payment
    - Debit card
    - Credit card
    - Discount cards

This list is not complete, but it should give an idea of where and how this great amount of data is generated. Finding the right data sources is a crucial part for companies. The most commonly used data sources are electronic transactions, website logs and sensor information. Data sources need to fit the question that should be answered by the Big Data analysis. Generally, the sources can be subdivided into external and internal data sources.[101]

By reviewing the above stated data sources, everybody will recognize that individuals unconsciously feed the Big Data net, whereas the benefit lies with companies and elsewhere. By reaching a critical stage of digitalization, every private and career activity will leave digital traps.[102]

[99] https://www.facebook.com/
[100] https://twitter.com/
[101] Cf. Ohlhorst (2013), p.38f.
[102] Cf. Bachmann/Kemper/Gerzer (2014), p.22

This process of data generation, data processing and data usage is crucial for the understanding of the concept of Big Data and is well illustrated by the Big Data Primary Circuit, shown in Figure 3.1. Because of the usage of technical products, communication technology and online services, everyone gets part of this data generation process. By means of the gathered data, companies are able to develop their products perfectly fit to customers' needs and their behavior. This constitutes a major source of product innovation. Through the purchase and use of these new innovative products, with ever more sensors, more data is generated. This is the main-cause for or the todays exponential data growth.[103]

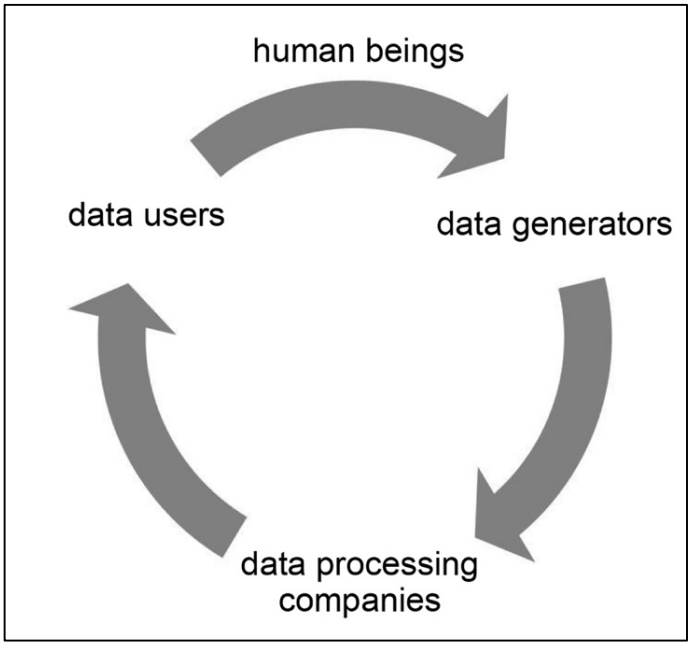


Figure 3.1: Big Data Primary Circuit[104]

3.3 Big Data Technical

On a high level of abstraction, with a focus on the technical side of Big Data, it can also be seen as the combination of hardware, methodic and mathematic. At first sight, this explanation does not appear to be new, since it is a basic characteristic of general data processing. Nevertheless, Big Data require a special development of these three topics and, additionally, not before all three areas are combined it can be called Big Data.[105]

Generally, databases underlie a certain scheme. These schemes define how the data needs to be stored in order to be processed and queried. A common scheme is the

[103] Cf. Bachmann/Kemper/Gerzer (2014), p.23
[104] Cf. Bachmann/Kemper/Gerzer (2014), p.23
[105] Cf. Bachmann/Kemper/Gerzer (2014), p.29f.

SQL (Structure Query Language) scheme. This common scheme unfortunately does no longer work with Big Data amounts. Hence, the data needs to be arranged in a different scheme. This new scheme is called NoSQL (Not only SQL) and forms a non-relational database. These databases form the basis for Big Data warehouses.[106]

For Big Data, two special types of hardware are needed, the first one being the hardware located at the companies. This hardware provides the computing power and storage to wield the huge amount of data generated. Generally, this hardware takes the form of databases, specially designed for real-time handling of Big Data. But, not to be forgotten, the data sources (or generators, shown in section 3.2) are also a form of hardware needed to provide a company's infrastructure with the needed information. This hardware, in everyday use (like smartphones), are playing and will play a big role in the future development of Big Data.[107]

The second item necessary for processing Big Data are methods, which form a very abstract term. The basic meaning of Big Data methods is customized algorithms that form knowledge out of the raw data. Because of the requirements arising from the 5 'Vs', these algorithms are highly advanced and customized for the individual field of application. The last point necessary for Big Data systems is mathematics, referring to the high dependence on statistical and stochastic models.[108]

The Big Data software is highly dependent on the field of application. Additionally, it is a fast changing filed of research because of new methods and principles being developed on a regular basis. Since this chapter is intended to give an overview, confer the references for further readings on big data software.[109]

3.4 Big Data Tools in a Company's Framework

Traditional methods, in terms of business intelligence and data warehousing, focus on company data, which is generated by time through different activities of the company. The aim of these processes is a precise model of the past and the present. Big Data analytics offer the advantage of enabling predictive forecasting.[110]

Therefore, analyses are not further limited to the past and present, but extended to the future. In this case, a distinction can be made between:[111]

---

[106] Cf. Fasel/Meier (2016), p.377
[107] Cf. Bachmann/Kemper/Gerzer (2014), p.30
[108] Cf. Bachmann/Kemper/Gerzer (2014), p.30
[109] Cf. Schmarzo, (2013), p.180f.
[110] Cf. Bachmann/Kemper/Gerzer (2014), p.161f.
[111] Cf. Bachmann/Kemper/Gerzer (2014), p.162

- Descriptive analytics: for example, end-of-month adjustment derived from preexisting financial data.
- Predictive analytics: for example, sales forecasting, a potential company development according to existing data and even automated generation of suggested courses of action.

Generally, a trend can be seen, reaching from the description of the present and the past (descriptive analytics) to statements on the near future (predictive analytics), culminating in prescriptive analytics, where descriptions about the medium and long-term future are made. [112]

This shift in focus on the concept of time within data analytics is illustrated in Figure 3.2.

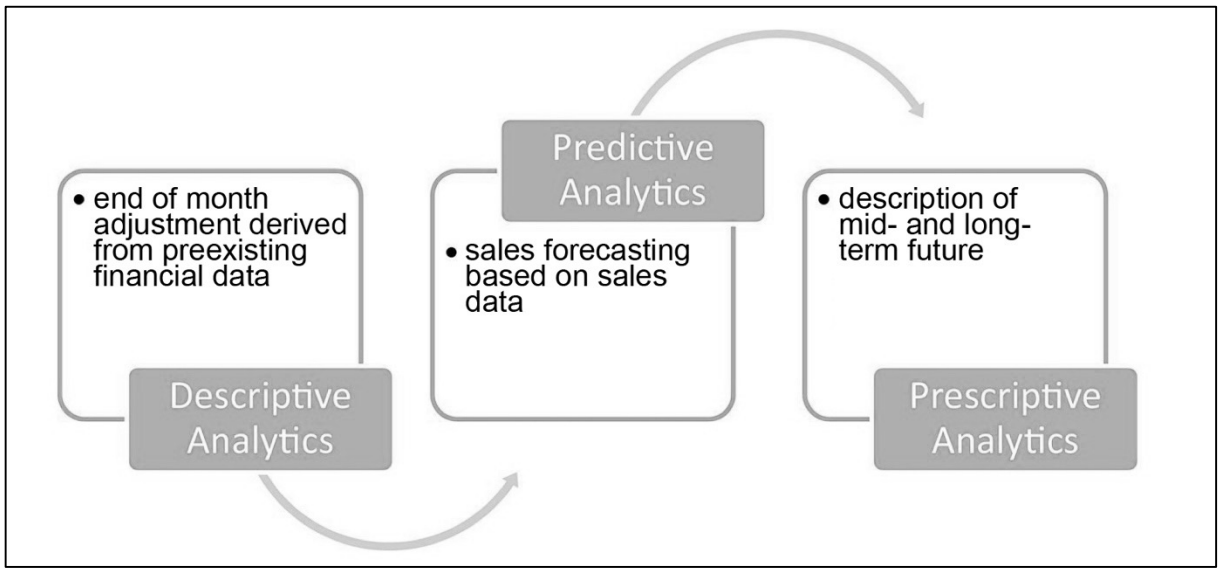


Figure 3.2: Evolution of data analytics according to the time horizon[113]

Big Data can be seen as a disruptive technology, not only because of the new hardware and software needed for the implementation, but moreover because of the necessity of implementing it for further business success. This setup brings many chief technology officers, chief information officers, and IT managers in the unpleasant situation to prove that Big Data brings value to their companies. Different tools, with different effort of implementation offer a smoother transition to Big Data analytics.[114]

[112] Cf. Bachmann/Kemper/Gerzer (2014), p.163
[113] Cf. Bachmann/Kemper/Gerzer (2014), p.164
[114] Cf. Ohlhorst (2013); p.21ff.

The depth of analysis varies considerably, starting from simple reports, where data is only conditioned and put into graphs to data mining processes. The latter one is of special interest for the marketing department, where data mining combined with stochastic methods is used to find patterns within the unstructured data, with the aim of identifying new customer needs and markets. Big Data analytics and data mining are similar to each other and get mixed up easily.[115]

3.4.1 Data Mining

Data mining is an already well established method in companies for the analysis and information generation from datasets. Since these datasets are generally big enough to play a role for Big Data, the information retrieval of data mining is based on statistical models and usually deals only with structured or semi-structured data. In the last few years, methods from the field of artificial intelligence are also diffusing into the data mining process. Generally, two different types of analysis can be distinguished: [116]

1. Traditional hypothesis-driven analytics: The traditional hypothesis-driven analysis postulates a working hypothesis that determines which data is processed with what problem statement. The aim is to either verify or falsify the given hypothesis.
2. Data-driven analytics: Data mining can be categorized as a data-driven analysis. Data mining is an automated process for the advanced search of knowledge in datasets, by means of analyzing and describing found patterns. Therefore, data mining can be seen as explorative. With this approach, no hypothesis or model is preset.

Data mining is a process where datasets are analyzed from different perspectives in order to generate a manually reviewable summary. Generally, data mining uses data at rest or archival data. Methods used in data mining focus on modeling and knowledge discovery for predictive, rather than descriptive, purposes. It forms an ideal process for uncovering new patterns from large data sets.[117]

There are different application classes for data mining, which can be seen in Table 3.1. For this thesis, the class of text mining is especially interesting, thus this class will be discussed in section 4.

---

[115] Cf. Bachmann/Kemper/Gerzer (2014), p.162
[116] Cf. Bachmann/Kemper/Gerzer (2014), p.164
[117] Cf. Ohlhorst (2013), p.4

For data mining, expert knowledge that is often underestimated is needed because of the absence of a hypothesis and the explorative behavior. They form a high potential of misinterpretations. The question of who is in charge of interpreting the data mining results in a company is not trivial and needs to be answered carefully. Additionally, the runtime of data mining algorithms needs to be sufficiently short. These fast algorithms often lack in statistical accuracy, which brings in additional uncertainty about the results.[118]

Class	***Task***	***Application***	***Example of method***
Association	Identify and quantify dependencies and relations	Market analysis	Statistical dependency analysis
Clustering	Identify groups of objects on the basis of similarities	Customer segmentation	K-means algorithm
Classification	Allocate objects within already defined classes	Solvency check	Rule induction
Text mining	Derive structured data from unstructured content	Web mining and information retrieval	Search algorithms and linguistic methods
Forecasting	Calculation of future parameters according to independent variables	Churn prevention	Regression and neuronal networks

Table 3.1: Application classes for data mining[119]

3.4.2 Predictive Analytics

Data analyses, which enable forecasts for any more volatile market, get more and more important for a company's strategic decision making. Statements influencing key factors, like turnover or profit generated from deep knowledge of customers and markets, form a major part of predictive analytics. By means of these forecasts,

[118] Cf. Bachmann/Kemper/Gerzer (2014), p.171
[119] Cf. Bachmann/Kemper/Gerzer (2014), p.166f.

companies can better act on the fast-changing markets and better address different customer groups.[120]

Implementing predictive analytic tools into key business processes will lead to predictive answers of certain key questions, like:[121]

1. Where to optimize network operations, marketing spending, and staffing decisions?
2. What's the financial impact of pricing, route, or supplier changes?
3. What's the business potential to score partners for quality, delivery, and service reliability?

To ensure significant forecasting results, the corresponding department (e.g. marketing) needs extensive communication with the implementing IT department. On the one hand, mathematical as well as statistical knowledge and, on the other hand, specific knowledge on the corresponding department's field of operation is needed in order to build applicable and correct predictive models.[122]

At first sight, verifying such a predictive model may seem an easy task; one needs to compare reality with the results of the predictive model. This is indeed easy, as long as the model proofs to be accurate. If this is not the case, finding the real mistake in the predictive model is not easy at all. For this reason, three aspects have to be taken into account for the iterative optimization of predictive models:[123]

1. Selection of relevant data
2. Quality of selected data
3. Quality of the predictive analytic model according to reality

Especially when dealing with unstructured textual data, the quality of the selected data strongly influences the results. Textual data can differ in many ways, like being differently encoded, for all sorts of websites. Additionally; they can be written in different languages and can vary widely in length, all effecting the processing algorithms. These problems are not as present with semi-structured and structured data.[124]

Each of those three factors strongly influences the results of a predictive model. Hence, measures to consider these factors influencing the results need to be developed.

---

[120] Cf. Bachmann/Kemper/Gerzer (2014), p.171f.
[121] Cf. Schmarzo (2013), p.84f.
[122] Cf. Bachmann/Kemper/Gerzer (2014), p.172
[123] Cf. Bachmann/Kemper/Gerzer (2014), p.173
[124] Cf. Ittoo/Minh Nguyen/van den Bosch (2015), p.103

Generally, these factors are iteratively generated by adapting one factor at a time and comparing the analysis results with reality. Thus, the influence of each factor can gradually be determined. [125]

Basically, Big Data analytics can be seen as a fusion of traditional methods like data mining and predictive analytics. The aim of Big Data analytics is not only to describe the blurred image of the short, medium and long-term future, but a sharp one, including recommended calls for action for either acting or counteracting different predicted chances and/or risks.[126]

In future, the implementation of analytic tools to forecast key questions within and outside the companies' environment, will be a key success factor for companies. The goal is to successfully implement analytics-driven business applications, such as category management and demand-based forecasting.[127]

3.4.3 Prescriptive Analytics

Prescriptive analytics is the third and last part of the superior class of business analytics:[128]

1. Descriptive analytics
2. Predictive analytics
3. Prescriptive analytics

By outbidding all technical opportunities and combining it with statistical and mathematical models, it becomes possible to predict future events. Figure 3.3 shows the enhancement of prescriptive analytics as an evolution of data mining and predictive analytics. For reasons of obviousness, the field of descriptive analytics is not displayed in the figure.[129]

Prescriptive analytics in contrast to predictive analytics is not focused on a single question that should be answered. Prescriptive analytics describes a future state, based on the results of behavioural analysis on already existing data as well as the results from real time-data processing.[130]

---

[125] Cf. Bachmann/Kemper/Gerzer (2014), p.173
[126] Cf. Bachmann/Kemper/Gerzer (2014), p.174
[127] Cf. Schmarzo (2013), p.24
[128] Cf. Bachmann/Kemper/Gerzer (2014), p.175
[129] Cf. Bachmann/Kemper/Gerzer (2014), p.175f.
[130] Cf. Ittoo/Minh Nguyen/van den Bosch (2015), p.103

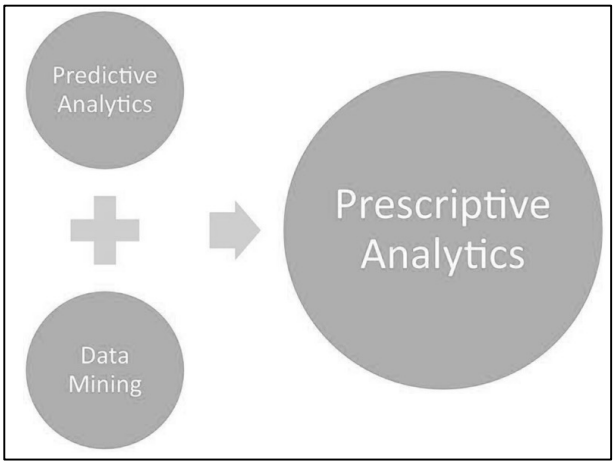


Figure 3.3: Prescriptive analytics as evolutional merge of predictive analytics and data mining[131]

Big Data Analytics, by means of behavioral analysis of already existing data in combination with real-time data processing (InMemory), enables prediction of future events. Four main constraints have to be observed, to ensure plausible results for the Big Data analytics[132]:[133]

1. Big Data: The dataset to be analyzed needs to be big enough to bring sufficient diversity with it. In other words, the dataset needs to be big enough to not be strongly influenced by either side-effects or individual abnormalities.
2. Data Mining: By means of mathematical as well as statistical and stochastic models, patterns in the datasets are found. According to these patterns, probabilities are calculated for certain scenarios happening in the future leading to new developments.
3. Predictive Analytics: The forecasts are calculated for different future timespans.
4. InMemory: InMemory is necessary due to the technical requirements of real-time data processing. This is a technological alternative to the traditional hub and spoke network architectures. Highly simplified InMemory does no longer write the data to storage, but simply keeps it in a computer's main-memory. For further details on InMemory, confer the references.

[131] Cf. Bachmann/Kemper/Gerzer (2014), p.175
[132] Cf. Bachmann/Kemper/Gerzer (2014), p.176
[133] Cf. Bachmann/Kemper/Gerzer (2014), p.176f.

4 Text Mining

Text mining is part of the superior class of data mining. The literature does not offer a clear definition of data mining. It is frequently confounded with terms such as knowledge discovery, machine learning and predictive analytics, each of which have a slightly different meanings depending on the context. A definition of data mining found in the literature is the screening of existing data in order to find useful patterns that can be interpreted and further processed.[134]

Data mining can be divided into two approaches, depending on the format of the existing data to be analyzed. On the one hand, data mining can refer to the analysis of structured (most numeric or categorical) data or, on the other hand, the analysis of unstructured data, commonly text. The latter is referred to as text mining as a subgroup of data mining.[135]

A definition for text mining fitting the superior class of data mining is offered by Bhanuse et al., which says that:

"Text Mining is knowledge discovery process from large database to find out unknown patterns."[136]

A wide variety of toolsets is offered for the analysis of structured data. Due to efficiency, a big effort prevails to use this already developed and well known toolsets also for text mining purposes. For this reason, the text mining process, described in the next chapter, processes the unstructured textual data into structured data sets, where common toolsets can be applied.[137]

Hence, the data gathered by the screen routine, described in section 5.2.1, is mainly plain text; the subsequent sections focus on the basic principles of text mining only.

4.1 Text Mining Process

The goal of text mining is to convert unstructured text into semi-structured data. Because many the tools of data mining can be applied to semi-structured data, like clustering, classifying or predicting, this is not possible with unstructured data.

[134] Cf. Vijay/Balachandre (2015), p.2
[135] Cf. Vijay/Balachandre (2015), p.275
[136] Bhanuse/Kamble/Kakde (2015), p.807
[137] Cf. Ester/Sander (2000), p.274

Additionally, this semi-structured data base is the basis for machine-learning algorithms, which can be trained to understand the semantics of given texts.[138]

The text clustering, categorization and summarization systems formed a central research topic within the field of information retrieval. While structured data is managed with database systems, text data is typically managed via search engines. The search engines offer the user to find useful information from a collection conveniently with a keyword query, also used in the later developed screen routine.[139]

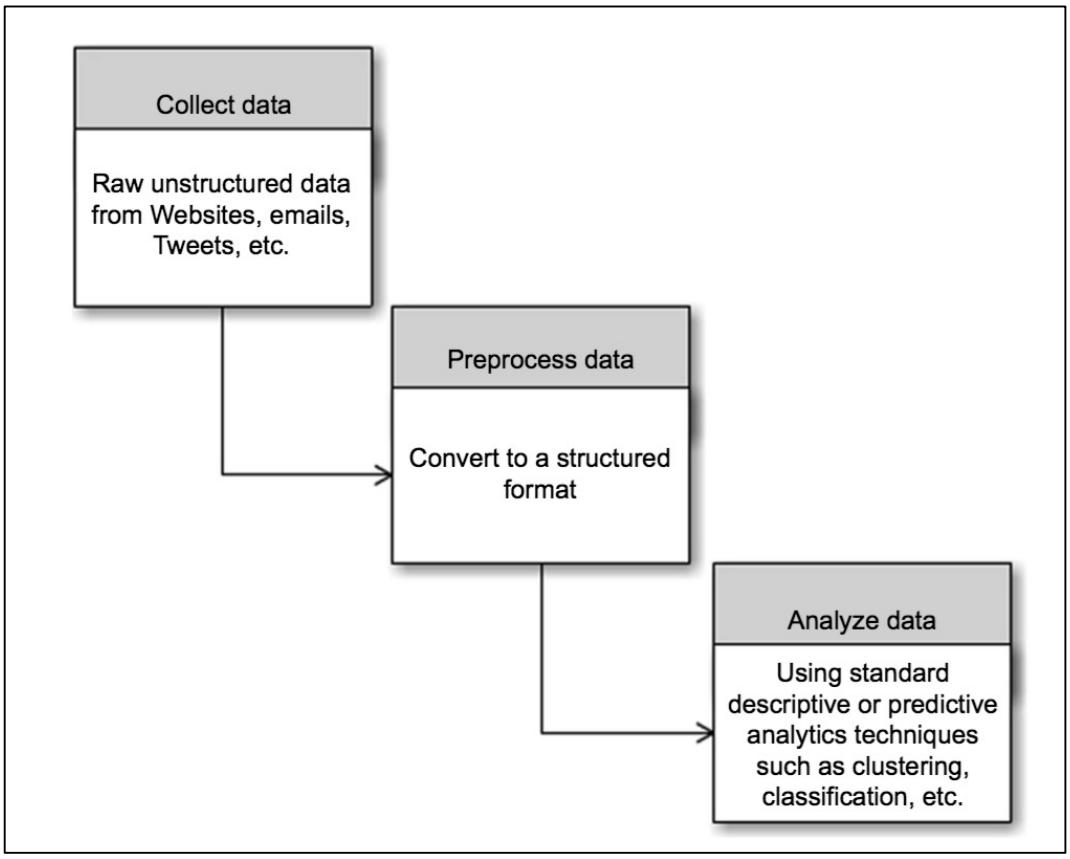


Figure 4.1: High-level abstraction of the text mining process[140]

Figure 4.1 shows the three main steps of text mining. Each of these steps is described in the following chapters.

[138] Cf. Vijay/Balachandre (2015), p.277
[139] Cf. Aggarwal/Zhai (2012), p.2
[140] Cf. Vijay/Balachandre (2015), p.277

4.1.1 Collecting Data

According to the literature and Figure 4.1, the first step of text mining is collecting data. Since this process equals the information sourcing activity of the technology forecasting process, the same tools and frameworks are used here. Hence, for information on collecting data for text mining, see chapter 2.3.2.

4.1.2 Preprocessing Data

Data collection is followed by preprocessing of textual data. Here, unstructured text is converted into a semi-structured dataset. For this purpose, a set of text mining tools is available. These tools can be arranged individually to best fit the given task.[141]

Generally, a distinction between different text mining approaches can be made. For both approaches the preprocessing is a necessary step before the data processing can start. The two different approaches differ in their use of external data. In classic text mining tools, no additional information on the text is gathered, only the words are analyzed. That is different with eg. Natural Language Processing (NLP). There the algorithms make use of external sources, sometimes linked with the text, to better derive the overall concept of the text's sentiment.[142]

Before starting with preprocessing itself, the quality of the textual data needs to be checked. Therefore, simple spellchecking is required. Additionally, special characters need to be removed before starting with the initial preprocessing. The last thing to ensure good results is to change the case, either putting into upper case or lower case all documents. This means that all words in all documents are either in upper case or lower case letters. Having imposed these limitations, the preprocessing can start.[143]

Step	*Action*	*Result*
1	Tokenize	Convert each word or term in a document into a distinct attribute
2	Sopword removal	Remove highly common grammatical tokens/words
3	Filtering	Remove other very common tokens
4	Stemming	Trim each token to its most essential minimum
5	n-grams	Combine commonly occurring token pairs or tuples

Table 4.1: Typical sequence for preprocessing tools used in text mining[144]

[141] Cf. Vijay/Balachandre (2015), p.283
[142] Cf. Ester/Sander (2000), p.276
[143] Cf. Vijay/Balachandre (2015), p.283
[144] Vijay/Balachandre (2015), p.283

In Table 4.1, a typical sequence of preprocessing tools can be seen. By referring to this table, the single tools are described in the next subchapters.

4.1.2.1 Tokenization

As first step of text preprocessing, the text itself needs to be segmented. Generally, the text can be separated into words, phrases, sentences or abstracts. For standard text mining processes history showed that dividing the text into it's single words sets the best basement for further analysis.[145]

Fortunately, languages show some common behavior, one of which is that words are separated by blanks. This special character can be used to separate the words of a document. Each word of the document is then called a token, and the process is called tokenization.[146]

A more precise definition of the word token is given by Rafael E. Banchs:

"A token is defined as any string that can be captured by using the grouping operator within a given regular expression."[147]

Tokenization is the first step of bringing text into a semi-structured format. For this, a matrix is built, the term-document matrix. This is a binary matrix with as many rows as tokens in the documents and as many columns as documents in the example set. For each token in a given document, the matrix entry equals one. If a token does not occur, the entry equals zero.[148]

For better understanding, a small example of two documents, consisting of just one sentence each, is shown. The documents can be seen in Table 4.2. The results of the tokenization can be seen in Table 4.3. This is the term-document matrix.

Document 1	This is a chapter on text mining.
Document 2	This chapter describes the basics on text mining.

Table 4.2: Example-set consisting of two documents with one sentence each

[145] Cf. Carstensen et al. (2010), p.264
[146] Cf. Vijay/Balachandre (2015), p.279
[147] Banchs (2013), p.42
[148] Cf. Banchs (2013), p175

	a	*basic*	*chapter*	*describes*	*is*	*mining*	*on*	*text*	*the*	*this*
Document1	1	0	1	0	1	1	1	1	0	1
Document2	0	1	1	1	0	1	1	1	1	0

Table 4.3: Term document matrix resulting of the example-set

In this matrix, as a result of tokenization, the unstructured text has been transformed into a structured matrix. This is essential for many further analyzing tools. Most of them rely on the structured term document matrix and it constitutes a cornerstone for machine learning algorithms.[149]

4.1.2.2 Stopword-Removal and Filtering

By having a closer look at the examples from Table 4.2, one may recognize the frequent occurrence of words such as ‘a’, ‘this’, ‘and’ and similar words. In common documents consisting of thousands of words, these frequent words make up the majority of the words in the documents. Therefore, grammatical words like articles, conjunctions, prepositions, and pronouns need to be filtered before further text processing is carried out. This process of removing common words without adding information to the text is called stopword removal and the words themselves are thus called stopwords.[150]

To illustrate this, a diagram can be seen in Figure 4.2. It shows Zipf’s law, which states that a very small number of words forms the largest proportion of a written text. For this diagram, the sixty-six books of the Bible, with more than 30 000 verses, is analyzed. The plot shows the word frequency over its rank, where the rank indicates the ranking of a word according to the number of occurrences (e.g. rank 1 means that this word occurs most frequently).[151]

Additionally, two more findings are made by George K. Zipf. He found out that the most frequently used words are also the shortest ones and the less used words tend to get longer with increasing rank. Secondly he mentioned that with lower rank the semantic significance of a word reduces. Saying, that a word frequently used contributes less to the text’s sentiment than a less frequently used one.[152]

---

[149] Cf. Vijay/Balachandre (2015), p.280
[150] Cf. Vijay/Balachandre (2015), p.280
[151] Cf. Banchs (2013), p.114
[152] Cf. Hausser (2000), p.322f.

It can be noted that the plot shows a linear behavior in the middle section of ranks and slightly deviates at both extremes. Especially the first three ranks constitute a special situation. These are the words ‘the’, ‘and’ and ‘of’ and therefore the three most frequent words used in the Bible. By calculating the sum of the frequencies of these three words, the total amount they occur, is 150233 times. By taking the seven highest ranked words and sum up their occurrences, one will see that they make up more than 25 % of all the words written in the Bible; an overview is given in Table 4.4.[153]

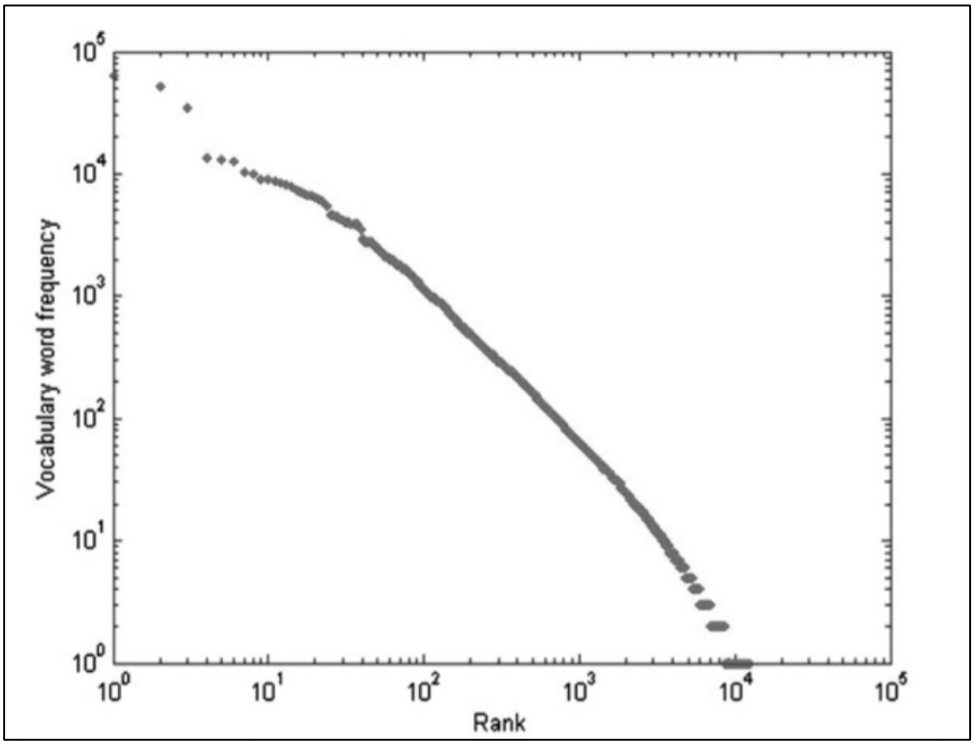


Figure 4.2: Plot of word frequencies versus rank to illustrate Zipf’s Law[154]

Rank interval	***1-7***	***1-42***	***1-249***	***1-1147***
Percentage of vocabulary	0.06%	0.33%	1.98%	9.14%
Total number of occurrences	199795	395047	593882	712352
Percentage of whole collection	25.25%	49.92%	75.04%	90.01%

Table 4.4: Clarification of Zipf’s Law via the n most frequent words[155]

[153] Cf. Banchs (2013), p.117f.
[154] Banchs (2013), p.117
[155] Banchs (2013), p.118

Stopword filtering is normally directly performed after tokenization. For instance, standard English stopwords were removed from the example in Table 4.5. As one can see, the dimension of the matrix is drastically reduced by simply filtering the standard grammatical words. The size of the matrix halved, from 10 columns to 5, and the semantics of the text can still be recognized. Keeping in mind that normal documents consist of thousands of words; therefore, the reduction of the matrix's size is essential to keep the needed computing power as low as possible.[156]

	basic	***chapter***	***describes***	***mining***	***text***
Document1	0	1	0	1	1
Document2	1	1	1	1	1

Table 4.5: Term document matrix of the example set after stop word filtering

In addition to filtering standard English stopwords, other specific terms can also be filtered. For example, dealing with the automotive industry, it is helpful to filter out common words in that industry, such as 'car', 'automobile', 'vehicle', and so on. This is commonly achieved by creating a custom-made dictionary containing all the words to be filtered. [157]

Apart from custom-made or standard stopword dictionaries, there is a second possibility to find stopwords. By simply analyzing the frequency words occur at in a document. Standard English stopwords will be filtered because they occur frequently and, on the other hand, very specific terms will be removed because they occur very rarely. Bearing in mind that exactly these specific words could be the ones characterizing the text. Nether the less, this method has the advantage that there is no need for the creation of stopword dictionaries.[158]

4.1.2.3 Stemming

When writing grammatically correct texts, words with similar contextual meaning very often need to be conjugated or declined, which leads to small modifications of their spelling. For example, the word recognize must be modified to recognized, recognizable, recognition and so on according to grammatical function they fulfil in a clause. But all of these modifications have the same grammatical root 'recognize'. Stemming means reducing words to their root or stem. This process simplifies the

[156] Cf. Vijay/Balachandre (2015), p.282
[157] Cf. Vijay/Balachandre (2015), p.282
[158] Cf. Banchs (2013), p.205f.

conversion of unstructured text to structured data because now, only the root terms occur in the term document matrix. Thus, all words with the same contextual meaning are merged into one word, the word stem.[159]

The most common stemming technique for English texts is the Porter method. This technique works based on several rules with the basic idea of removing and/or replacing the suffix of words. One rule, for example, is to replace all words that end in 'ies' by 'y' (e.g. companies turns into company). Similarly, a second rule is to stem all terms ending in 's' by removing the s (e.g. algorithms turn into algorithm). The Porter technique is simple and easy to apply, but makes mistakes. For example, according to the above rules, both words 'arms' and 'army' are stemmed to 'arm'. Not representing the original meaning of the initial word. In addition to the Porter method, other stemming methods exist. It depends on the applied domain and experience which one to choose.[160]

4.1.2.4 N-Gram Modeling

N-gram modeling is a common tool in basic text mining. It is a statistical model, that avoids creating tools describing the sentences' grammar. N-gram models can't give as precise results on single words significance according to the sentence sentiment as a grammar model could. Nether the less, because of their easy and universal implementation, they form a cheap alternative to understanding the grammar underlying in a given document.[161]

There are phrases used in spoken and written text that typically go together. They are referred to as collocations in linguistics. For example, the word 'good', is very frequently followed directly by either 'morning', 'afternoon', 'evening' or 'night'. Collocations are called n-grams in the context of text mining. Looking at a text not in single words only, but in n-grams, can make a big difference.[162]

The basic principle behind n-gram models lies in the Markov probability. The Markov model describes the probability of random processes in which the probability of the next state only depends on the current state. For n-gram models, this means that the probability of sequences of certain words are calculated and, according to the Markov probability, n-grams are formed or not. Applied to the above example, the probability

[159] Cf. Vijay/Balachandre (2015), p.282
[160] Cf. Vijay/Balachandre (2015), p.282
[161] Cf. Indurkhya/Damerau (2010), p.343f.
[162] Cf. Vijay/Balachandre (2015), p.283

of ‘good morning’ to form a bigram is high, whereas the probability of, e.g. ‘good electricity’, is low and thus these instances are treated as two separate words.[163]

Search engines were the first algorithms to make use of such n-gram models on a large scale. They use the models for automatic translation, identifying speech patterns, checking misspellings, entity detection, information extraction, and many other applications. Most commonly used are bi- and trigrams, consisting of either two or three words. N-gram modeling is typically the final step in forming the structured term document matrix.[164]

4.1.3 Analyzing Data

Data analyzing is the last step in the text mining process. This is the step where the aimed information is generated. For this, different tools are available. The most common tools, like document categorization and document search, are described in the following chapters.

4.1.3.1 Document Categorization

For document categorization, the nature of a textual document is crucial, because according to this approach, the different documents are grouped together in different data collections, called classes or categories. In general, there are many useful applications of document categorization such as spam filtering, press clipping or document clustering, to just name a few. Alternatively, document clustering sets the first step for further, more complex analysis, like opinion mining and plagiarism detection.[165]

In essence, document categorization can happen via three different methods:[166]

1. unsupervised clustering
2. supervised classification in vector space
3. supervised classification in probability space

Human minds depend very strongly on classes and categories, collections of objects are always tried to be arranged in groups. Therefore, categorization processes are the basis for the abstraction process of human thinking. Such processes are oriented towards finding the most prominent attributes of the objects to find common characteristics, which in turn become the subjacent characteristic of the class itself. In

---

[163] Cf. Banchs (2013), p145f.
[164] Cf. Vijay/Balachandre (2015), p.283
[165] Cf. Banchs (2013) p.237
[166] Cf. Banchs (2013) p.237

terms of textual data, the most projective attributes are the words and their sentiment. Thus, organizing text documents into groups is a problem of a semantic nature.[167]

Unsupervised clustering automatically groups objects of a given collection according to likeness of their most projective attributes. Moreover, it is called unsupervised clustering because there is no information on the classes known beforehand, that is, the user does not predefine any classes at all, they are all generated automatically according to the most projective attributes of the documents.[168]

The most commonly used algorithm for unsupervised clustering is the k-means algorithm, as shown in Figure 4.3. It is an iterative algorithm, where a collection of n observations is partitioned into k clusters. This process is iteratively repeated until the set converges into a stable partition. Each iteration takes two steps. The assignment step and updating step. In the assignment step, each document is assigned to the class that best fits the most projective attributes. Hence, the classes themselves slightly change during each assignment step, because of new attributes joining the class. Thus, in the second step, the update step, the classes are updated and documents that no longer fit into the class, are rearranged. As already mentioned, this process is repeated until a stable condition is reached.[169]

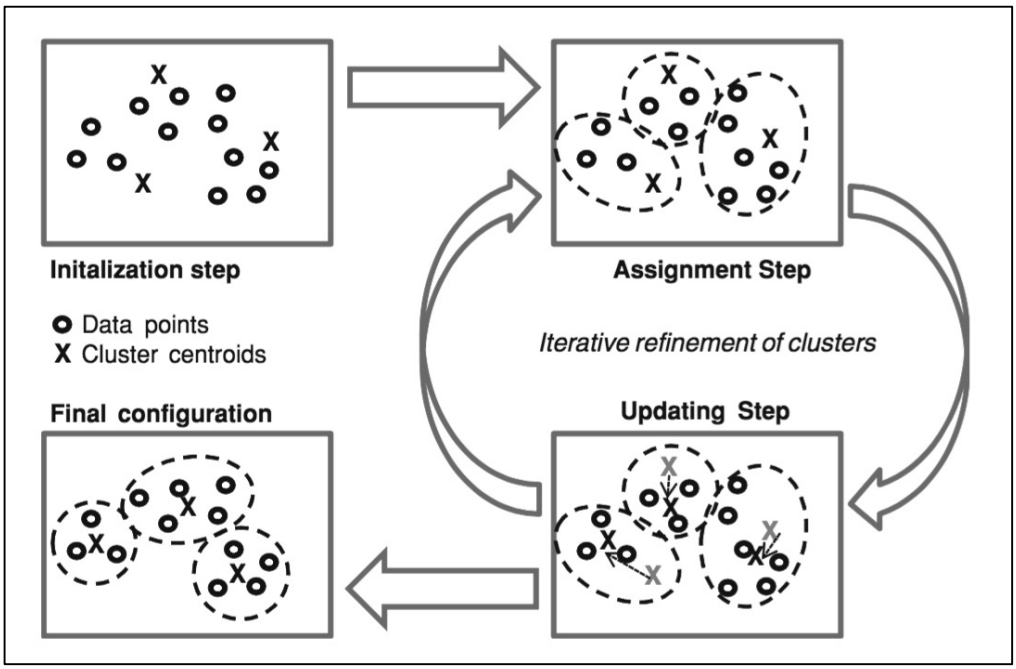


Figure 4.3: Illustrative example on the operation of the k-means clustering algorithm[170]

[167] Cf. Banchs (2013) p.242
[168] Cf. Banchs (2013) p.243
[169] Cf. Bhanuse/Kamble/Kakde (2015), p.809f.
[170] Banchs (2013), p.243

Supervised clustering, in contrast to unsupervised clustering, requires previous knowledge about the data collection. This knowledge can usually be gained through data samples, for which categories are already known. These data samples are important for the algorithms to learn which projective attributes are important for the different classes. This learning process is then used to categorize new data samples.[171]

In general, the initial information needed for supervised clustering comes from the text iteself or already available example texts. In the first case side information, so called meta-data, linked with the text offer the possibility to find inputs for supervised clustering algorithms. The second case gives algorithms the possibility to train with example texts, already containing the meta-data. This second possibility is used, if the initial text doesn't offer meta-data links right away.[172]

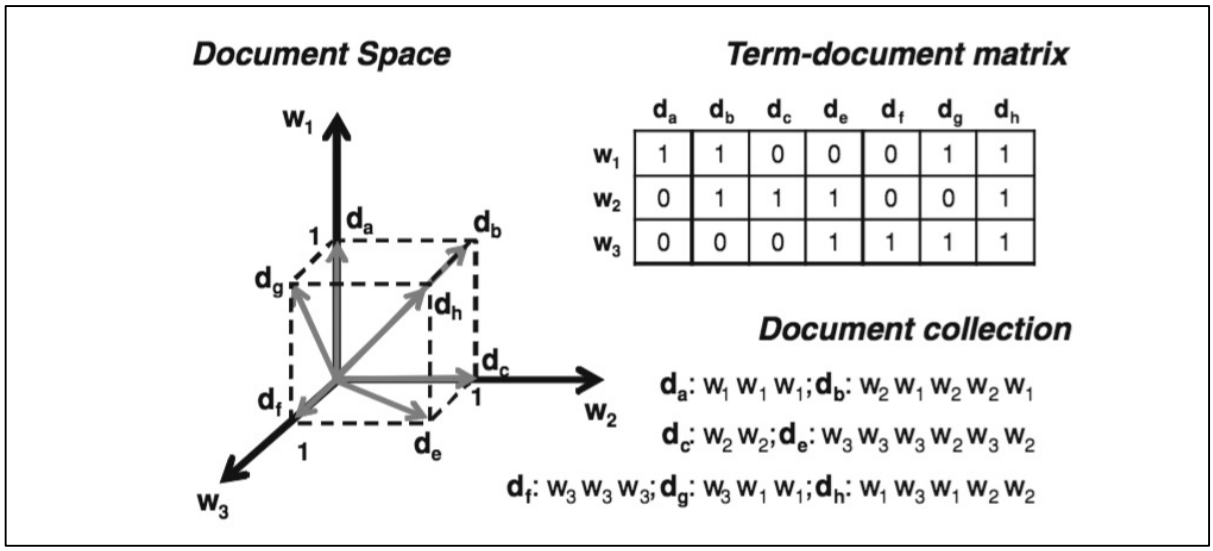
The most common algorithm for this process is called k-nearest-neighbour (knn). Similar to the k-mean algorithm, the knn-algorithm also works with a vector space model. For the categorization of a new sample, the knn-algorithm aims at the nearest sample for which a category is already known. The nearest sample in this context means the nearest sample according to the vector space model.[173]

We will take a look at a term-document matrix as the one computed in chapter 4.1.2.1. Now we look at the columns of the term-document matrix as vectors, representing documents over an n-dimensional vector space, with n as the total number of words. In Figure 4.4, a simple vector space can be seen, with a data collection of seven documents containing the three words w1, w2 and w3. As seen in the figure, each binary vector in the term-document matrix represents one document. Each vector has three components, which correspond to the three words w1, w2 and w3. Together they form an orthogonal basis in a three-dimensional vector space. In this simple example, following observations can be made[174]:[175]

1. One vector, representing a document consisting of only one word, is collinear to the corresponding word axis.
2. One vector, representing a document not containing one of the three example words, is orthogonal to the corresponding word axis.

[171] Cf. Banchs (2013), p.252
[172] Cf. Bhanuse/Kamble/Kakde (2015), p.810
[173] Cf. Banchs (2013), p.252
[174] Cf. Banchs (2013), p.183f.
[175] Cf. Banchs (2013), p.183

3. Two vectors, representing two documents with the same words just in different order, show the same vector representation.
4. Two vectors, representing two equal documents, show the same vector representation.



	d_a	d_b	d_c	d_e	d_f	d_g	d_h
w_1	1	1	0	0	0	1	1
w_2	0	1	1	1	0	0	1
w_3	0	0	0	1	1	1	1

Figure 4.4: Illustrative example of the document space for a sample document collection[176]

Now it is possible to define the phrase 'nearest neighbor'. For the supervised clustering in vector space, this means that the angle between two vectors (and therefore two documents) represents the similarity between these two. The more similar, the smaller the angle is, the less similar, the bigger. This will be further explained in chapter 4.1.3.2.[177]

Finally, there is the possibility of supervised classification in probability space. In contrast to the methods described before, which worked on a geometrical model, the underlying models of probability space classification are statistical models. For further information, confer the references.[178]

4.1.3.2 Document Search

Document search forms the key problem of information management, which is also relevant for text mining applications. The superior class of document search is the

[176] Banchs (2013), p.183
[177] Cf. Banchs (2013), p.183
[178] Cf. Aggarwal/Zhai, (2012), p.181f.

study of information retrieval. There are two main tools available for document search[179]:[180]

1. Binary search
2. Vector search (based on the vector space model, like described in the previous chapter)

Binary search represents the simplest form of document search. The boundary conditions for a binary search are keywords that need to be defined. Subsequently, the algorithm searches the selected documents according to the defined keywords. The results are binary, meaning they are either 'yes' or 'no' apply. There is no information available about the frequency of the keywords that occur. Nevertheless, the binary search is a powerful tool, especially when keywords are combined with logic conjunctions like 'and', 'or' and so on.[181]

For example, searching for a whole sentence or even a summary does not make sense, because if there is only a difference of one character, the binary search will not find it. Hence, the abstract, or the sentence, needs to be reduced to a number of keywords. By combining this keywords with logic conjunctions a smart binary search can be set up.[182]

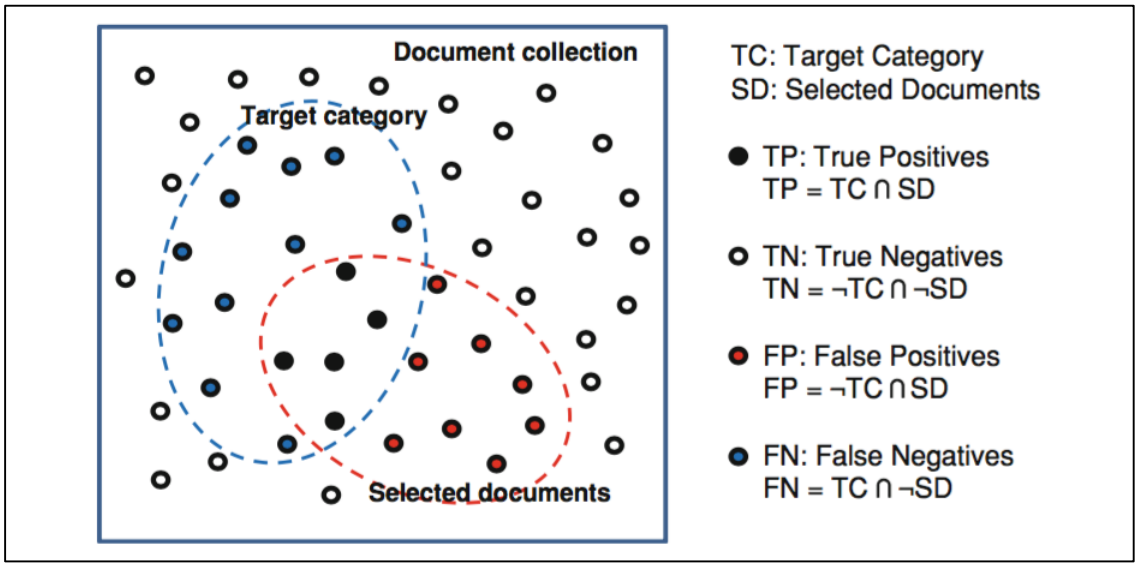


Figure 4.5: Target category and selected document subsets in a generic document search experiment[183]

[179] Cf. Ester/Sander, (2000), p.245f.
[180] Cf. Ester/Sander, (2000), p.245f.
[181] Cf. Aggarwal/Zhai, (2012), p.12f.
[182] Cf. Banchs (2013), p.278
[183] Banchs (2013), p.281

In Figure 4.5, the document collection can be seen. The target category represents the documents that should have been found by the binary search. The selected documents are the ones found by the binary search. From this setup, four subsets are formed:[184]

1. True positives (TP): This subset contains all documents correctly identified as belonging to the target category. True positives are the intersection between the target category and the selected documents.
2. True negatives (TN): This subset contains all documents correctly identified as not belonging to the target category. True negatives are the intersection between the complements of the target category and the selected documents.
3. False positives (FP): This subset contains the documents that were mistakenly identified as belonging to the target category. False positives are the intersection between the selected documents and the complement of the target category. They are also referred to as wrong selections.
4. False negatives (FN): This subset contains the documents that were mistakenly not identified as belonging to the target category. False negatives are the intersection between the target category and the complement of the selected documents. They are also referred to as missed selections.

Now getting back to the vector space model. Vector search is based on the vector space model described in the previous chapter and is an alternative to the binary search. This tool makes use of the vector space concepts of distance and similarity to define a search criterion that, in contrast to the binary search, produces a natural ranking mechanism for the selected documents. As opposed to the binary nature, the notion of similarity allows for introducing a continuous notion of relevance. This does not only open the possibility that documents are relevant or irrelevant with respect to a given query but another document can now be of more or less relevance than others.[185]

The result of a vector search is not a subset of the example set, it is a list of documents ranked according to their relevance relative to the query. This notion is very common, because this is exactly what modern search engines do (if payed advertisements are neglected).[186]

[184] Cf. Banchs (2013), p.281
[185] Cf. Aggarwal/Zhai, (2012), p.12f.
[186] Cf. Banchs (2013), p.290

4.1.3.3 Content Analysis

In the previous two chapters, we dealt with classification of documents as well as possibilities of searching documents according to defined keywords. This chapter continues in a similar direction, however from a different point of view. Now, the focus is shifted to subjective or opinionated aspects of documents rather than topics and specific domains. All of these aspects of a document can be seen as dimensions. There are several dimensions for documents, and working on basis of these is called dimensional analysis. Some other dimensions are[187]:[188]

1. Writing style or genre of the text
2. Authorship
3. Type of communicative function
4. Emotional and subjective characteristics

All of these dimensions can be analyzed conducting content analysis. For giving an overview one dimension will be treated here, the polarity distinction of opinion.

Polarity estimation is an important task for documents, especially when combined with intensity estimations. While polarity estimation gives binary results, with the two possibilities of positive or negative, intensity gives a continuous result by defining to what extent the text is either positive or negative. The intensity is generally given in percent. An example result could be, that the analyzed text is 87 % positive.[189]

Typically, the problem of polarity and intensity estimation is dealt with in a statistical way. The most common approach is the Naïve Bayes approach based on the likelihood ratio method. Since this approach is based on statistical models, a sufficiently big dataset is needed. This dataset needs to be analyzed beforehand, which means that the dataset of polarity as well as intensity has already been allocated to every single document. This dataset can then be used as a train-and-test dataset. With this train-and-test approach, it is ensured that the Bayes approach reaches accuracies of roughly 80 %, meaning that 80 % of the estimated polarities are correct.[190]

Figure 4.6 shows some examples of polarity and intensity estimation. Example number 6 is an interesting one. ‘Not so bad, I was expecting something worse’, implies that the event was not that bad, meaning the polarity should be slightly positive. But according

[187] Cf. Banchs (2013), p.313f.
[188] Cf. Banchs (2013), p.313f.
[189] Cf. Banchs (2013), p.319
[190] Cf. Aggarwal/Zhai, (2012), p.416

to the occurrence of the words ‘bad’ and ‘worse’, the polarity was estimated as negative and its intensity as -0,99, meaning absolutely negative. This example perfectly shows the limitation of such content analysis. It is human language that hides the real context within the sentences and which cannot be estimated by statistical methods. Example 8 can also be seen as a misinterpretation.[191]

	Example	Loglirat	Polarity	Score
1.	It was as good as garbage.	–2.7479	Negative	–0.8796
2.	This actor is terrible; his performance was pathetic.	–2.8371	Negative	–0.8893
3.	The music was bad and the script was boring.	–4.3772	Negative	–0.9752
4.	This film has some problems with the plot.	–1.0615	Negative	–0.4860
5.	Not as good as the previous movie in the saga.	2.0801	Positive	0.7779
6.	Not so bad. I was expecting something worst.	–5.0530	Negative	–0.9873
7.	The plot was simple, but I enjoyed the movie anyway.	–1.0728	Negative	–0.4903
8.	Interesting film, which is full of action and excitement.	–0.3436	Negative	–0.1701
9.	A very funny and pleasantly entertaining film.	0.7424	Positive	0.3550
10.	Wonderful script and beautiful photography. A great movie!	2.3287	Positive	0.8225
11.	Excellent movie, as expected from such a great director.	1.0857	Positive	0.4951
12.	Exceptional production I will be watching again and again.	1.3771	Positive	0.5971

Figure 4.6: Examples on polarity detection and intensity estimation[192]

Natural language processing (NLP), which will be described in the next chapter, tries to deal with this issue of interpreting human speech and word choice.

4.2 Natural Language Processing

Natural Language Processing (NLP) can be seen as the realization of natural language phenomena on computers. NLP is a field of computer linguistics, where informatics and linguistics meet. Methods of both fields are used in combination to overcome the challenge of processing human speech, either in written or spoken form, with computers. NLP had its early beginnings in the 1950s, but especially in the last few decades, with the occurrence of smartphones and other mobile devices, this field has advanced considerably.[193]

There are different knowledge scopes that are necessary for NLP, which are primarily defined through the different levels of description from linguistics. The following horizontal classification consist of five topics: [194]

[191] Cf. Banchs (2013), p.327
[192] Banchs (2013), p.327
[193] Cf. Carstensen et al. (2010), p.2
[194] Cf. Herbrich/Graepel (2010), p.4

1. Phonetics: Phonetics are of special interest when it comes to speech recognition. Phonetics deal with the articulation and characteristics of natural spoken language. One aim of phonetics is finding models that define certain attributes of sounds in words, for example, whether a vowel is voiced or unvoiced.
2. Morphology (or lexical analysis): Morphology deals with the formation and structure of words. Lexical analysis examines lexical roots of words and processes that affect the occurrence of certain word forms and, often, their meaning, for example, the 's' as a marker of the plural in English (paper–papers).
3. Syntax (or syntactic analysis): Everything defining the structure of a sentence is part of syntax. This is a key part of NLP, since the recognition of a sentence's structure is crucial for the further grammatical definition, and consequently the sentence's meaning. Hence, syntactic analysis not only recognizes grammatical deviations, but it also explains the relation of words in a sentence and recognizes the underlying structure.
4. Semantics: Semantics deal with the meaning of words, sentences and phrases. There are two parts. First, the meaning of lexical entities is described and then the meaning of bigger constructions like sentences, paragraphs, texts is examined.
5. Pragmatics: Pragmatics concentrate on statements and their purpose in the context they are uttered. For example, have a look at the statement 'The window is open!'. This utterance can have various meanings. The most likely one is that the person saying this is cold and s/he wants to have the window closed. In this context, the statement needs to be classified as a request to close the window.

The processes shown above are summed up in Figure 4.7, in the form of a flowchart.

Information Retrieval (IR) is one of the success stories in NLP. Since the Internet's content expands exponentially, and with the emergence of search engines like Google[195], NLP has been playing a major role in IR. According to Indurkhya & Damerau, IR can be defined as:[196]

[195] https://www.google.com/
[196] Cf. Indurkhya/Damerau (2010), p.455

"IR deals with the representation, storage, organization of, and access to information items. These information items could be references to real documents, documents themselves, or even single paragraphs, as well as Web pages, spoken documents, images, pictures, music, video, etc."[197]

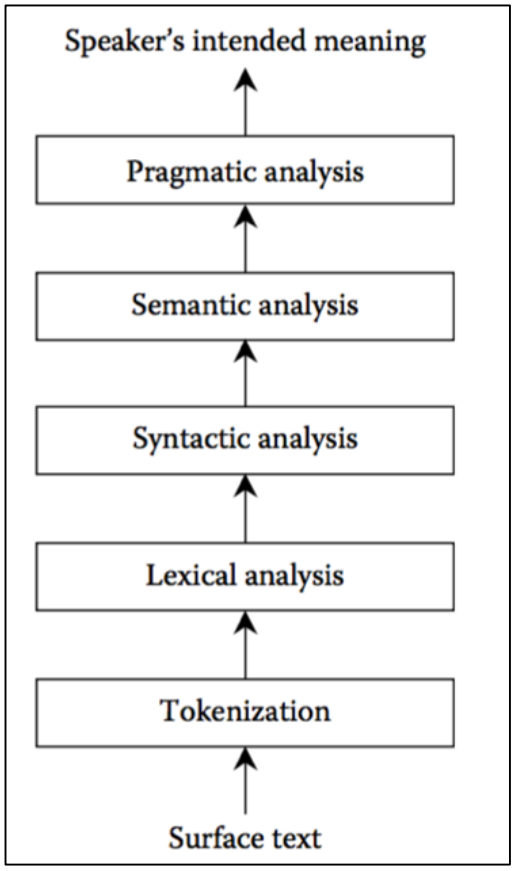


Figure 4.7: The stages of analysis in processing natural language[198]

IR has to deal with imprecise definitions on both sides, the user and document side. Because users often search for things they cannot articulate or they simply do not know what the exact term they are searching for is, IR has to derive concepts of this query that fit the search. The same thing happens within documents, where words are inflected or simply synonyms are used. Nevertheless, search engines need to find this information and match the search defined imprecisely with the document written imprecisely. Here, the word imprecise refers to the typically used queries in databases, where only 100 % matching strings or numbers can be found.[199]

---

[197] Indurkhya/Damerau (2010), p.455
[198] Herbrich/Graepel (2010), p.4
[199] Cf. Indurkhya/Damerau (2010), p.455

Apart from IR, NLP offers a variety of other applications:[200]

1. Correction systems: For example, the auto-correction function every smartphone supports. When texting on a smartphone and there are typing errors in the text, the smartphone automatically corrects them.
2. Computer lexicography and terminology: This is a still semi-manual job, consisting in feeding lexica with new terms to improve the NLP algorithms.
3. Text-based information management: This is the superior class of IR. There are other subtopics like text summarizing or question answering.
4. Speech recognition and generation: It is also well-known thanks to smartphones. 'Hello Google'[201] and 'Siri'[202] offer both perfect speech recognition and generation.
5. Machine translation: Machine translation, like Google Translator[203], also forms part of the NLP toolkit.

[200] Carstensen et al. (2010), p.14f.
[201] https://www.google.com/
[202] http://www.apple.com/de/ios/siri/
[203] https://translate.google.com/

5 Development of the Screen Routine

As described in chapter 1.1, efficient and easy-to-use tools play a major role for future decision-making.

The aim of this thesis is to develop a screen routine and demonstrate its feasibility. This approach should form a base for further research to build on. For this reason, some simplifications as well as some boundary conditions are set:

1. Amount of implemented data sources: For the demonstration of properties and the general feasibility of the approach, only one data source is implemented in the screen routine. Science Direct[204] as a quality source with a lower time lag than e.g. patent data bases is the one chosen for implementation.
2. Fully automatic data processing: The screen routine is not programmed for full automatic use yet. The approach builds on an agglomerate of code pieces written in MatlabR2015a[205], PythonIII[206] and IBM Bluemix[207]. The code was written to proof feasibility but is not a major part of this thesis. Nevertheless, the code can be found in the appendix, and is referred to in the corresponding chapters.

The development objective of this routine is to make it easy and intuitive to operate for the user. It should be applicable not only to strategic decisions. It should help employees in all different working areas like R&D, Mergers & Acquisitions (M&A), competitor analysis and venture capital investments.

To do so, literature has been studied, as well as a small sample experiences has been gathered to identify important characteristics. In the following chapter, these experiences are shared and a brief overview of the routine is given.

5.1 Holistic Framework of Embedded Screen Routine

For better understanding, the process of identifying key characteristics for the success of the developed screen routine is shown. Bearing in mind that these characteristics make comprehending certain setups, structures and desired properties of the developed screen easier. Additionally, a brief overview is given on the routine as a whole.

---

[204] http://www.sciencedirect.com/
[205] http://de.mathworks.com/
[206] https://www.python.org/
[207] https://www.ibm.com/cloud-computing/bluemix/

5.1.1 Identification of Key-Characteristics for a Successful Screen Routine

Before starting with the initial development of the screen routine, uncertainties about the key-characteristics and components of the screen routine need to be resolved. First, literature on different forecasting, data mining, and technology management topics is reviewed. Details on the single domains can be found in chapters 2, 3 and 4.

Nevertheless, a short summary of the identified key characteristics for a successful screen routine is stated below:

1. Time-lag: Depending on the branch a company operates in, time-lag can be crucial. For instance, the telecommunication branch, one of the most volatile industries, is dependent on fast forecasting tools. The state-of-the-art forecasting method used are technology scouts, because of their low time-lag resulting from their independency of publications.[208]
2. Individual uncertainty: Scouts and experts always rely on the opinion and behavior of single individuals. Hence, there is no possibility to guarantee a structured approach to the forecasting topic, neither can impartiality be guaranteed.
3. Source quality and quantity: To prevent forecast results from being one-sided, a high quantity of data sources needs to be ensured. Simultaneously, the quality of the data sources needs to be monitored, in order to not fall victim to misinformation.
4. Importance of expert opinions: Even though experts and scouts can be biased, the importance of them is not to be neglected. This is because they are able to do what no algorithm can do: Interpreting weak signals and logically combine pieces of information according to the topic's relevance.

After having defined a guideline for important characteristics via reviewing literature, a field experiment has been conducted to vanish last ambiguities. The Graz University of Technology (TU Graz) offers the necessary preconditions and familiarity to conduct such an experiment.

In the beginning, the structure of the TU Graz so that, it can be examined to find possible data sources from which information on technologies under development can be retrieved. This is illustrated in Figure 5.1. In blue, the TU Graz and its different organizational units are shown whereas in green, possible sources for information retrieval are represented.

[208] Cf. Rohrbeck (2006), p.1

The TU Graz consists of seven faculties, each of which consists of different institutes. Every institute has different fields of expertise. The blue bracket on the left side indicates that, with the TU Graz library and Pure, all faculties with all their institutes can be covered. The 'Forschungs- & Technologie-Haus'[209] (F&T-Haus) deals with the exploitation of TU Graz know-how generated internally, especially by means of patents. The online system of the TU-Graz (TUG-Online[210]) provides access to publications as well as graduate theses from all faculties and institutes.

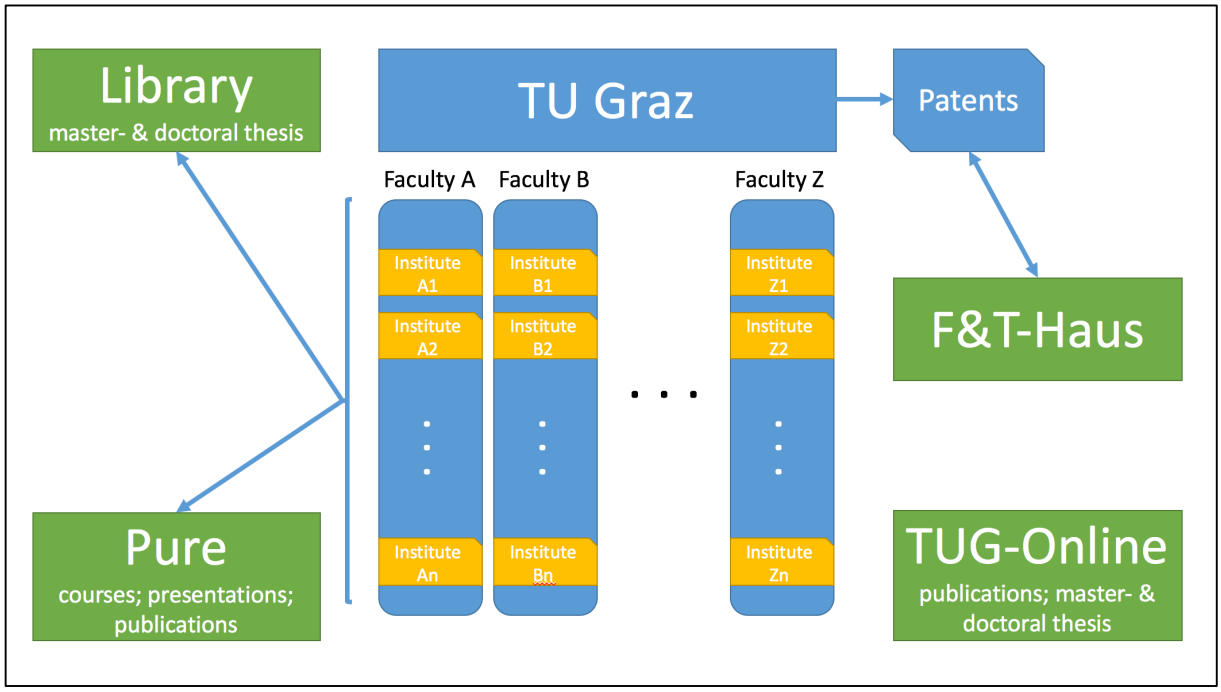


Figure 5.1: Possible information sources within the TU Graz's framework

These four sources are compared based on their properties. This can be seen in Table 5.1. By comparing the advantages and disadvantages of these sources, the TUG-Online[211] proves to be the most promising source of input for a screen routine.

Having defined a possible source, a customized downloader (similar to the Science Direct[212] downloader described in chapter 5.6.1) is coded. With this downloader, abstracts of master- and doctoral thesis can be downloaded automatically.

[209] https://www.tugraz.at/tu-graz/organisationsstruktur/serviceeinrichtungen-und-stabsstellen/forschungs-technologie-haus/
[210] https://online.tugraz.at/tug_online/
[211] https://online.tugraz.at/tug_online/
[212] http://www.sciencedirect.com/

Source	*Useful information offered*	*Access*	*Notes*
Library of TU Graz	master- & doctoral thesis	login	good for finding specialized theses, bad for an overview
Pure[213]	publications, fair participations, presentations, ...	unre-stricted	perfect for getting a fast overview, unfortunately, no abstracts or summaries of the activities and articles are available
TUG-Online[214]	master- & doctoral thesis; projects and research areas	unre-stricted	well-structured for searching all thesis within a research area, hard to find a specific thesis
F&T-Haus[215]	technologies offered for third parties	unre-stricted	the detailed list is very useful with all the information given about the different technologies, unfortunately, the list seems to not be up to date

Table 5.1: Comparison of information sources within the TU Graz's framework

Technologies from selected startups, with a clear link to the TU Graz, are determined. The aim is to find early indicators for these startup technologies among the downloaded thesis.

It is not possible to identify a direct connection between a specific startup technology to any thesis of the TU Graz, at least not before reaching a very high abstraction level. The following example will clarify the findings.

Reactive Reality[216] is a startup from Graz working with augmented reality. Their product Pic2Fit[217] is an app for mobile devices that lets you try on clothes virtually. In this case, the identified technology of this company is augmented clothing, virtual clothing, or similar. Important for the company's technology is the fact that it combines trying on clothes with augmented reality.

[213] https://pure.tugraz.at/portal/
[214] https://online.tugraz.at/tug_online/
[215] https://www.tugraz.at/tu-graz/organisationsstruktur/serviceeinrichtungen-und-stabsstellen/forschungs-technologie-haus/
[216] http://www.reactivereality.com/
[217] http://www.reactivereality.com/pictofit/

By scanning the TU Graz thesis, a great volume of research can be found in the field of augmented reality, which, on a high abstraction level, is Reactive Reality's[218] technology. But there was not one thesis marking a weak sign for combining clothing with augmented reality.

The learnings from the experiment within the TU Graz environment are as follows:

1. Regionality: Information sources must not be limited to smaller regions, as the TU Graz is limited geographically, to a country or at least a continental, regarding the majority of its research.
2. Data compression: Even by the implementation of only one data source, thousands of thesis are downloaded, resulting in an enormous amount of data being analyzed and reviewed. Finding an appropriate way to compress this data into a manually reviewable amount is crucial.
3. Search direction: In the experiment, the search direction is given by means of the selected startups being examined. Additionally, the routine needs to be steered into the right direction by means of relevant input.
4. Manual selection: Even with only three startups having been looked at, the data content diverged significantly. The distinction of whether an abstract corresponds to the targeted topic or not depends only on nuances, resulting in a final step of manual selection of the preprocessed dataset.

According to these learnings, both from literature and the field experiment, several desired properties for the developed screen routine are derived. These properties can be seen in chapter 5.2.

5.1.2 Overview of the Developed Screen Routine Approach

The screen routine can be used in many different fields like R&D, competitor analysis or venture capital marketing.

In chapter 5.4, the screen routine process is described in detail, but for better understanding a overview is given. In Figure 5.2, an illustration of the screen routine can be seen. It consists of four major parts:

1. Initial keyword generation: As the field experiment shows, a certain search direction needs to be set. To start the search and steer it into the right direction, an initial set of keywords needs to be defined. Generally, initial keyword

[218] http://www.reactivereality.com/

generation is up to the user, but as part of this research, a way of semi-automated keyword definition is found for a general trend analysis.

2. Data gathering within selected sources: After the keywords are defined, the data sources are screened for relevant information on the given keywords. All information found, that corresponds to the keywords' topic is downloaded.
3. Natural Language Processing (NLP) of gathered data: As seen in the key characteristics, data compression is important to enable the final manual review, for which NLP is chosen. Through the data gathering process, huge amounts of textual data are collected. This text is not manually reviewable in an appropriate time-span. Therefore, the different articles are analyzed using a NLP algorithm. This algorithm compresses the information into new keywords, concepts and entities.
4. Steering the screen routine: Steering the routine is a result of several key characteristics like search direction, expert opinion or final manual review. The list with the new keywords, concepts and entities is revised by the routine's operator. By reviewing the concepts, entities and keywords, the user can define which words to use for the next search round and which ones to put into the technology catalog for further analysis.

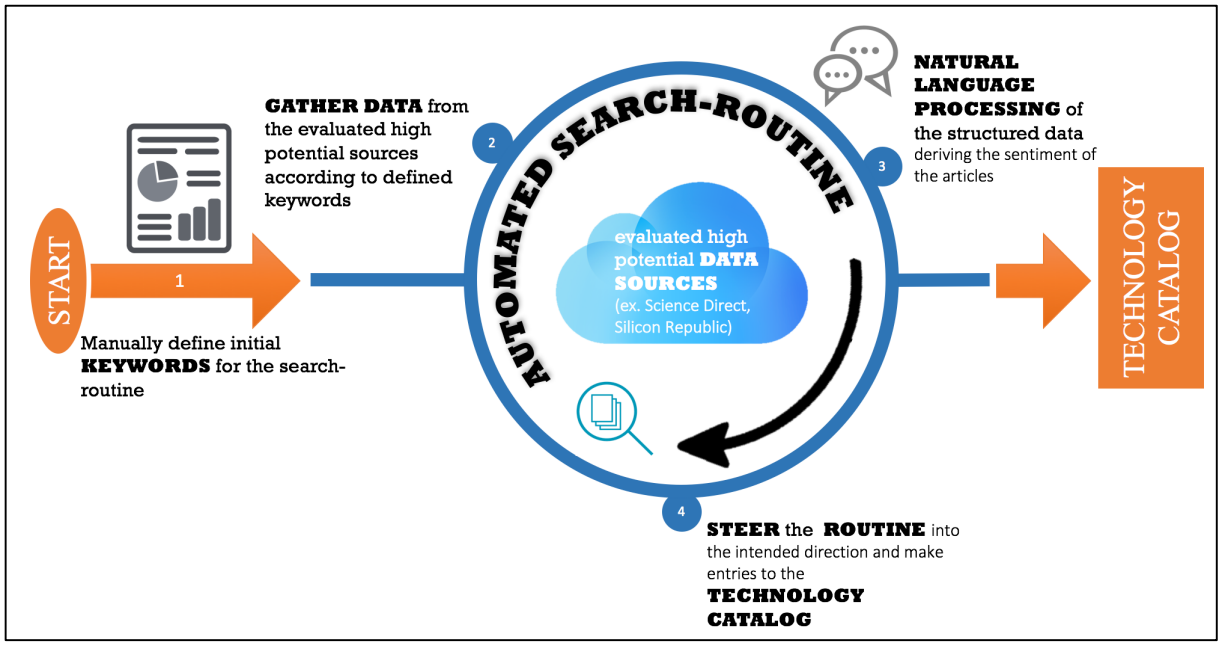


Figure 5.2:Illustration of the screen routine approach

The output of the screen routine is a technology catalog. Upon completion of each search round, the operator adds relevant technologies and trends to the technology catalog. Hence, with each search round, the catalog grows. Thus, this technology

catalog is the ideal basis for later decision-making or analyzing processes. An example of such technology catalog can be seen in chapter 5.5.

5.2 Desired Properties and Requirements

The fields of application for the developed screen routine are broad. To guide the development process, some properties and requirements of the routine are defined beforehand. The approach of literature review and a field experiment described in chapter 5.1.1 formed the base for the formulation of the desired properties and requirements. Some of these properties are shown in chapter 5.6, where examples of the test routine are given.

The following properties and requirements were defined:

1. Data gathering: The screen routine needs to collect huge amounts of data from different kinds of sources so it will not be one-sided.
2. Data distillation: This is a key requirement for the routine's success. It needs to compress huge amounts of data into a manually reviewable data set, without losing too much information.
3. Routine maneuverability: It must be possible for the user to steer the routine in the intended direction. For example, it must be possible for the operator to either dig deeper into a single topic or, if this is not wanted, to wander off to different topics.
4. Easy and intuitive operation: Like the data distillation, this is also a key requirement. Since the routine is developed to be used in a broad field of application, it must be easy and intuitive to use.
5. Routine speed: In order to make use of the routine on a frequent basis, it is not sufficient to have an easy and intuitive operation only; the process should only take an appropriate amount of time to deliver results.

5.2.1 Data Gathering

The results of a process are only as good as the input. Consequently, the data gathering sets the foundation for reliable results of this screen routine. If data gathering was the only requirement to be fulfilled, the easiest way to gather data would be to download every website possible. Since there are other properties influencing the routine's performance, a compromise needs to be found.

There are two extreme scenarios for this case; the one of downloading a very small amount of information relevant to the topic and, on the other side downloading everything that can be brought into context with the topic in one way or another.

The first one bears the danger of getting one-sided results. If the amount of data downloaded is too small, the results are dependent on this small amount of information. For example, if the blog entries by a single author were the only input to the screen routine, then all the results delivered by the routine would be dependent on this one individual opinion. This is not useful for the routine's performance, thus, it must be ensured that an adequately set of data is gathered.

The second extreme is gathering all information that can be brought into relation with the topic. Due to the fast growth of the unstructured data available online this process is a mammoth task and not feasible with the developed approach. Another interesting factor is the reliability of online sources. Because many websites are free to use, e.g. social media, blogs and so on, users can easily become authors. Everybody can state his/her opinion online without the content being reviewed. This kind of information is not beneficial for this routine either.

Somewhere between these two extremes, a compromise for the application of this routine can be found. In summary, these are the important points gained:

1. Number of information sources: In order to avoid one-sided results, the number of information sources and authors needs to be larger than a critical amount.
2. Routine performance: To avoid making the routine becoming inert, the number of information sources must not be too big either. A good compromise for the number of sources needs to be found.
3. Reliability of information sources: To avoid false information, the reliability of the sources needs to be checked.

In conclusion, the number of information sources as well as the type are crucial to the screen routine's success. Further details can be found in chapter 5.3.

As mentioned above, data distillation is one of the key requirements for satisfactory operation of the screen routine. Without efficient data compression, the routine would be worthless.

In the end, it is the operator who decides which information is relevant and which not. Thus, the final data distillation, or categorization (in useful and not useful technologies), still needs to be done manually, especially because the screen routine should be applicable to different fields of application. Therefore, the automated part of the routine needs to compress the data to a certain amount, possible to be reviewed manually.

Like the data sources, the compression of information can be done in different ways. The kind of data used for the routine is mainly unstructured data, more precisely,

textual data. Therefore, the only efficient way to handle this kind of information is by text mining. This process and especially NLP are described in chapter 4.2.

If the search routine gathers thousands of documents, each of which consist of thousands of words. Keeping in mind that the result of the routine is a list of keywords and entities of a considerable amount and which will be reviewed manually, a problem is faced. This problem is a standard text mining problem. Thus, text mining tools are applied to the documents from the data gathering. First, the documents are pre-processed to turn the unstructured textual data into a semi-structured format. This format enables further NLP algorithms to retrieve important information. The NLP algorithm derives concepts, keywords and entities fitting the sentiment of the original document. This is the only process ensuring that the raw data is compressed to an amount of keywords that is manually reviewable, without losing too much relevant information.

From there on, the operator manually edits the list that is the result of the fully automated part of the screen routine. This manual part of the routine can also be seen as steering the routine, which is described in the next chapter.

5.2.2 Routine Maneuverability

As stated in the introduction, the routine should fit for different fields of application. Thus, the routine needs to be manually steered into the intended direction. The result of the data compression process is a list that is manually reviewable. This is the starting point of the steering process of the routine.

The screen routine delivers information with every search round, which implies that the routine is iteratively steered into the intended direction. The search field (e.g. automotive industry) of the routine cannot be predefined, as one of the requirements is the broad use of the routine. Therefore, the search field is defined by a set of keywords, generated by the operator. After the first search round, the keyword set will deviate from the initial one.

The list delivered by the data distillation contains keywords, entities and concepts of the content that was found. Now, the operator reviews this list and decides in which direction to steer the routine. Depending on the operator’s aim, the following scenarios can unfold:

1. The operator is not satisfied with the generated list: Because the concepts and entities on the list generated by the data compression process does not satisfy the operator’s needs, he needs to adapt the initial keyword set for the next search round. By reviewing the list, the operator will find keywords that do not

fit the initial search intention, but the operator might nevertheless be interested in some of them. For this reason, the operator will include these keywords in the keyword set for the next search round and the results will show a tendency towards a new search direction. This will let the screen routine wander around in different search fields. This process offers high potential to identify white spots, as described in chapter 2.2.

2. The operator is satisfied with the generated list: The concepts and keywords on the list fit the operator's intentions. Thus, s/he will include some concepts and entities in the technology catalog, where s/he collects technologies and trends relevant to his/her purpose. Nevertheless, the operator might want to know more about certain concepts of the first search round, so s/he simply takes keywords that fit the topic and that s/he wants to know more about and includes them in the keyword set for the next search round. This round will generate a list with more detailed information on the chosen topics. As a result, the routine will focus on intended topics.

As can be seen in the two scenarios, the routine can be used either to wander around different topics, or dig deeper into one given task. The latter one is perfectly suited for solving determined problems. This maneuvering can be changed for every search round, resulting in an agile routine. The operator can decide in which direction to go very fast, offering the freedom to be as flexible as possible. Entries in the technology catalog ensure that no important information is lost during this screening process.

5.2.3 Easy and Intuitive Operation

To fulfil the requirements stated in chapter 1.2, the routine needs to be operable for employees working in different departments. Hence the operation needs to be simple and intuitive enough, so that the need for experts can be avoided.

The semi-automatization of the screen routine ensures easy operation. Every process requiring expert knowledge is automated. For example, the information compression, the decision of which information is kept and which is neglected is made by the NLP in a first step. The second step of information compression is conducted by the operator. Consequently, s/he decides which information of an already pre-filtered list is relevant for his/her task or not.

As a result, the operator needs to be well informed about the field of application the search routine is used in. However, s/he does not have to be specialist in computer linguistics, data mining or general analysis. Thanks to this characteristic, the routine can be used easily on a broad basis by any firm.

The steering process designed for this routine ensures intuitive operation. It is an intuitive behavior that the operator uses interesting keywords for the next search round. As a result, s/he can either examine different topics or merely focus one. This is something that happens intuitively and unconsciously.

5.2.4 Routine Speed

Because the routine is developed to be an iterative process, steered into the intended direction step by step, each iteration process needs to be sufficiently fast. If, for example, one search round takes a week to be finished, the whole screening process, then, might take, depending on the numbers of iterations, one or more months. This is too long for fast changing markets.

Thus, one requirement in its development is that one search round is finished within a maximum of a few hours so that a whole screening process could be done in less than a week. As a result, the screening can be conducted on a regular basis, depending on the branch it is used in and the change rates on a monthly or quarterly basis.

In addition to the problem of fast changing markets, the resulting costs of a slow screening process are unneglectable as well. If the company has to hire an employee especially for operating the screening routine, this results in expenses. Hence the routine is developed to be operated on a regular basis in parallel to the everyday work of the employee.

5.3 Selection of Data Sources

The starting point for every screening, forecasting and evaluation process is data sourcing. The quality and characteristics of data sources are crucial for the upcoming processes. Different sources show different characteristics. For example, very reliable sources tend to be inert compared to social media, where content is published in almost real time. One major goal of the data source selection process is to find a set of sources that blend together the different properties, in order to overcome drawbacks of single sources.[219]

By means of a literature research and studying Science Park Graz[220] startups, different sources suitable for this topic were identified, with various characteristics. Given that the developed screen approach automatically processes huge amounts of data, only online sources were of relevance for this thesis.

---

[219] Lin, Hu, & Wu, 2014, p.1
[220] http://sciencepark.at/

In a next step, those sources showing the same properties are categorized into clusters. Subsequently, these clusters are evaluated and combined to form an optimum data-source blend.

5.3.1 Single Source Evaluation

Before starting to evaluate data sources, a list of suitable sources needs to be generated. Literature research combined with studying startups and the platforms they use to communicate their progress, yielded the following list of suitable online sources. The list can be seen in Appendix 10.1. This list is not complete, but it provides a good starting point for further research.

These sources are compared in a value benefit analysis in order to get a rank of the sources with the highest potential for the developed screen routine. The value benefit analysis was exercised according to VDI2225[221].

Before the different properties of the sources are compared, criteria and their weighting need to be defined for the value benefit analysis. This was done by comparing the criteria pairwise, like shown in Figure 5.3. The following criteria were defined to identify sources that best fit the developed screen routine:

1. Time-lag: Describes the timespan between the publication of information and the point at which the information becomes accessible to the website personnel. This means that the time lag mainly describes the review process.
2. Frequency: Describes the frequency of new data being uploaded and available on a website.
3. Regionality: As shown by the field experiment in chapter 5.1.1, regionality is an essential characteristic of a data source. It describes the effective range of a website. If a website only operates locally within city boundaries, it is of less interest than websites operating globally.
4. Amount of relevant data: Describes the amount of data that is useful for the screened topic. For example, websites dealing with the intended search direction will offer more useful information than general news or social media websites.
5. Number of authors: Describes how many authors actively add content to the website.

[221] Norm 2225 des Verein Deutscher Ingenieure (1998)

6. Reliability: Describes the quality of information offered on the website, which is mainly defined by the process of publication. Sources are more reliable if the content is reviewed, and the more often or the higher qualified the personal reviewing, the higher the reliability. For example, patent websites are a very reliable source because of multiple expert reviews, in contrast to Facebook[222], which is an unreliable source because everybody is able to post anything without any reviewing process.
7. Ease of data gathering: Describes the difficulty of downloading data from a website automatically. Because if the data is not accessible via an automated algorithm, the best information source is useless for this screen routine.

		...more (1), equal (0.5) or less important (0) than...							
	Comparison of Criteria	Time-Lag	Frequency	Reginality	Amount of Relevant	Number of Authors	Reliability	Ease of Data Gather	Sum
Relating to criterion X is...	Time-Lag		0,5	1	0,5	1	0,5	1	4,5
	Frequency	0,5		1	0,5	1	1	1	5
	Reginality	0	0		0,5	0,5	0	0	1
	Amount of Relevant Data	0,5	0,5	0,5		0,5	1	1	4
	Number of Authors	0	0	0,5	0,5		0,5	1	2,5
	Reliability	0,5	0	1	0	0,5		1	3
	Ease of Data Gathering	0	0	1	0	0	0		1

Figure 5.3: Pairwise comparison of evaluation criteria for the value benefit analysis according to VDI2225[223]

According to the pairwise comparison, the following weightings (W) are calculated according to equation 1. Table 5.2 gives the parameters and the corresponding description for the calculation of the criteria weightings. As one can see in Figure 5.4, frequency and time lag are the dominating criteria. This is a result of the short time span in which a new technology brings value in case of a first-mover advantage. The ease of data gathering is prioritized lowly, which results in professional web crawling

[222] https://www.facebook.com/
[223] Norm 2225 des Verein Deutscher Ingenieure (1998)

tools, or even customized coded donwloaders enabling nearly any content to be downloaded.

Parameter	*Description*
W [%]	Criteria weighting in percent
S [-]	Sum for criteria according to Figure 5.3
n [-]	Total number of criteria

Table 5.2: Parameter description for criteria weight calculation[224]

$$W[\%] = \frac{s_j}{\sum_{j=1}^{n} s_j} * 100 \qquad (1)$$[225]

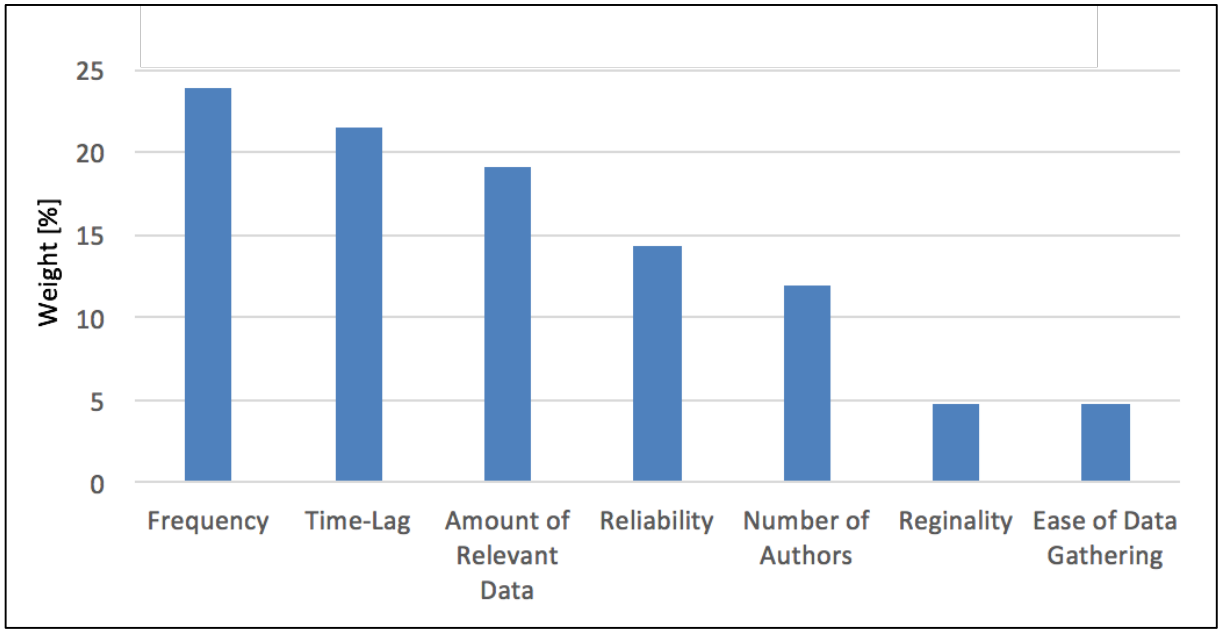


Figure 5.4: The weights of the different criteria as result of the pairwise comparison

According to VDI2225[226], the different sources now need to be evaluated on a scale of zero (no fullfilment of requirements) to four (complete fulfilment of requirements). Table 5.3 shows the score definition of the criteria.

[224] Norm 2225 des Verein Deutscher Ingenieure (1998)
[225] Norm 2225 des Verein Deutscher Ingenieure (1998)
[226] Norm 2225 des Verein Deutscher Ingenieure (1998)

	Score				
	0	**1**	**2**	**3**	**4**
Time lag	>12 months	6-12 months	1-6 months	1-4 weeks	1-7 days
Frequency	> once a year	once a year	monthly	weekly	daily
Regionality	< than city limit	city limit	state-limit	continent limit	global
Amount of relevant data	no relevant data	few data on everything	lots of data on everything	few but specific data	lots of specific data
Number of Authors	1 author	2-5 authors	6-20 authors	20-100 authors	>100 authors
Reliability	no review	single review	multiple review	expert review	multiple expert review
Ease of data gathering	no downloading possible	custom downloader	standard web downloader	query	newsletter

Table 5.3: Score definition of the criteria for the value benefit analysis

The list of sources can be evaluated in the value benefit analysis. This analysis can be seen in Appendix 10.2. The result is a list of sources ranked according to their potential use for the screen routine. The fifteen best ranked sources can be seen in Table 5.4.

	Source	***Points in the value benefit analysis***
1	Angel.co	336
2	Crunchbase	333
3	Science Direct	333
4	Scopus	333
5	Silicon Republic	333
6	Web of Science	333
7	Espacenet	331
8	Kickstarter	329
9	EPO	326
10	USPto	326
11	Interesting Engineering	321
12	Techbrunch	321
13	Wonderful Engineering	321
14	GeekWire	317
15	BBC	314

Table 5.4: List of 15 high potential sources as result of the value benefit analysis

With this result, the single source evaluation ends. Because the screen routine does not only rely on one source, but on many, it makes sense to have a closer look at the properties of different source categories. This is done in the next chapter.

5.3.2 Categorization of Sources

The categorization of sources was carried out in two steps. The first step was to define clusters of websites that share similar content. For example, there are different websites for patent data bases, but they all share similar content. Social media websites also share similar content and there are others apart from Facebook[227], Twitter[228] and so on.

To depict similar characteristics of the websites the scores in the value benefit analysis are compared. Websites showing similar scores are combined to a cluster, through this the following 8 clusters are formed:

1. Accelerators & Incubators (Acc/Inc): Science Park Graz[229], Indibio[230], …
2. Festivals & Fairs: TED & TEDx[231], Startup Grind[232], …
3. News & Techposts: Crunchbase[233], BBC[234], …
4. Patents & Trademarks: Espacenet[235], EPO[236], …
5. Scientific Publications: Science Direct[237], Scopus[238], …
6. Social Networks: Twitter[239], Facebook[240], …
7. Technology Reports: Boston Consulting[241], McKinsey[242], …
8. Universities: Massachusetts Institute of Technology (MIT)[243], TU Graz[244], …

---

[227] https://www.facebook.com/
[228] https://twitter.com/
[229] http://sciencepark.at/
[230] http://www.indiebio.co/
[231] https://www.ted.com/
[232] https://www.startupgrind.com/
[233] https://www.crunchbase.com/
[234] http://www.bbc.com/
[235] https://worldwide.espacenet.com/
[236] https://www.epo.org/
[237] http://www.sciencedirect.com/
[238] https://www.scopus.com/
[239] https://twitter.com/
[240] https://www.facebook.com/
[241] http://www.bcg.com/
[242] http://www.mckinsey.com/
[243] http://web.mit.edu/
[244] https://www.tugraz.at/home/

The result is eight different data clusters, all of which have different properties. The objective is now to combine some of these data sources in a way that the disadvantages of one website are balanced by the advantages of another.

To illustrate this process, a morphologic box[245] was created. The first column on the left shows the different criteria that were used in the value benefit analysis described in the single source evaluation. Next, in each criterion's row, the different possibilities of the extent to which this criterion is fulfilled are stated, starting with the worst possibility on the left (corresponding rating in the value benefit analysis 0) and the best possibility on the very right (corresponding rating in the value benefit analysis 4).

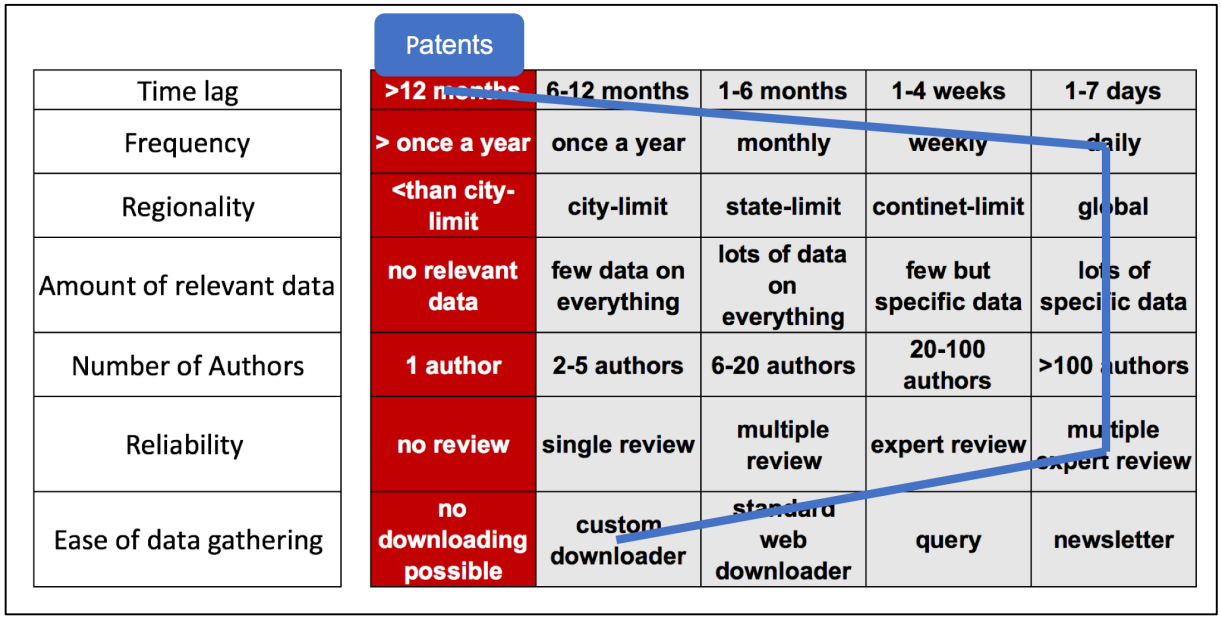


Figure 5.5: Morphologic box with illustrated properties of the patents & trademarks cluster

The zone marked in red, on the left side, indicates the area where sources should not be. Because, for example, having a time lag of more than one year is too long for the screening of new technology trends, except one is looking for already established technologies, but wants to reuse them in a new field of application. Therefore, this red zone does not mark a knock-out criteria, but, in general, data sources should be kept out of this zone.

Now the different source clusters can be added to this morphological box. For example, the blue line in Figure 5.5 represents the properties of the patents & trademarks cluster. It is interesting that, except from the time lag and ease of data gathering, patents & trademarks are great data sources according to the defined requirements. Thinking back to the value benefit analysis and the criteria weightings, time lag was the most

[245] Ritchey (2006).

problematic one. To adapt this property deficit of patents and trademarks, a data source that shows a very low time lag needs to be found.

In Figure 5.6, in addition to the patents & trademark cluster, the news & techpost cluster is added in orange. Having a closer look at the disadvantages of the patent & trademark cluster, the news & techpost cluster perfectly compensates this disadvantages. Repeating this process, the other clusters are also added to this morphological box.

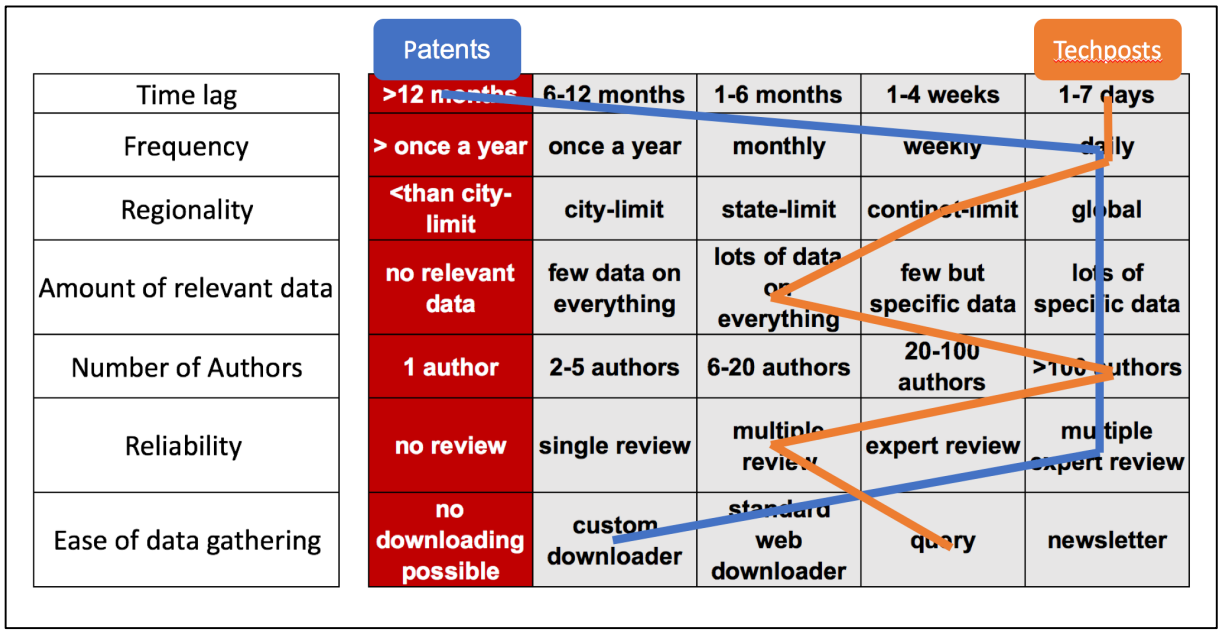


Figure 5.6: Morphologic box with illustrated properties of the patents & trademark cluster (blue) and the news & techpost cluster

Through this process, another interesting discovery has been made. By further grouping the clusters, different data source categories are found. By changing the perspective (of looking at the problem), three categories can be defined. Until now, the source evaluation process has concentrated on evaluating the generated data source list. Changing the way of looking at the screen routine as well as thinking about the ideal blend of data sources, delivers the following three categories:

1. Quality Sources (patents & trademarks, scientific publications, …): Generally, quality sources are defined through their high reliability gained due to multiple expert reviews. The resulting drawback is the time lag. In all other properties, these are very good data sources to retrieve information from, but not the newest information.
2. Dynamic Sources (news & techposts, social networks, …): Dynamic sources are the opposite of quality sources regarding the properties. They are defined by their short time lag and thus their low reliability. These sources are good for

obtaining the latest information and hyped technologies; nevertheless, the information needs to be treated carefully. An information validation is recommended, in order to not include misinformation for the screen routine.

3. Sophisticated Sources (accelerators & incubators, festivals & fairs, …): Sophisticated sources are defined by their content. They offer very specific information on the newest technologies and/or startups. From these sources, it is generally hard to retrieve information automatically, since sophisticated sources are very careful with online information. Nevertheless, this data source is perfectly suitable for steering and seeding information in the screen routine.

From these three sources, another criterion for data source selection has been found. The blend of data sources fed into the screen routine needs to have a balanced ratio of all three data source categories. This last criterion ensures that the screen routine operates on a broad information basis and does not concentrate on one field of application. For example, if the routine operated only on patent data, the results would be one-sided and dependent on the time lag, shifted in time. A balance of source categories ensures that disadvantages of sources are balanced by the advantages of other sources. For example, the high time lag of quality sources is balanced by the short time lag of dynamic sources.

In the next chapter, a small summary of the data source evaluation process is given.

5.3.3 Selected Data Sources

Before coming to the results, a small summary of the source evaluation process follows. First, possible data sources need to be gathered; for this, literature and, depending on the field of application, experience deliver numerous ideas for data sources. Then, criteria and their weighting for the evaluation of these sources are defined to further conduct a value benefit analysis. To avoid a one-sided search base for the screen routine, the sources need to be assigned to one of the three source categories (quality-, dynamic- or sophisticated source). Now, depending on the designated amount of implemented sources, an equal number of the best rated sources (according to the value benefit analysis) of each category need to be picked.

In this example, the fifteen best rated sources, depicted in Table 5.4, still need to be checked for the equilibrium of data source categories. Illustrated in Figure 5.7, the equilibrium of data source categories is not given at first (blue columns). To reach the equilibrium, dynamic sources need to be singled out and quality and sophisticated sources need to be included in the list. By doing so, the final result of the fifteen high potential sources to feed the screen routine can be seen in Table 5.5.

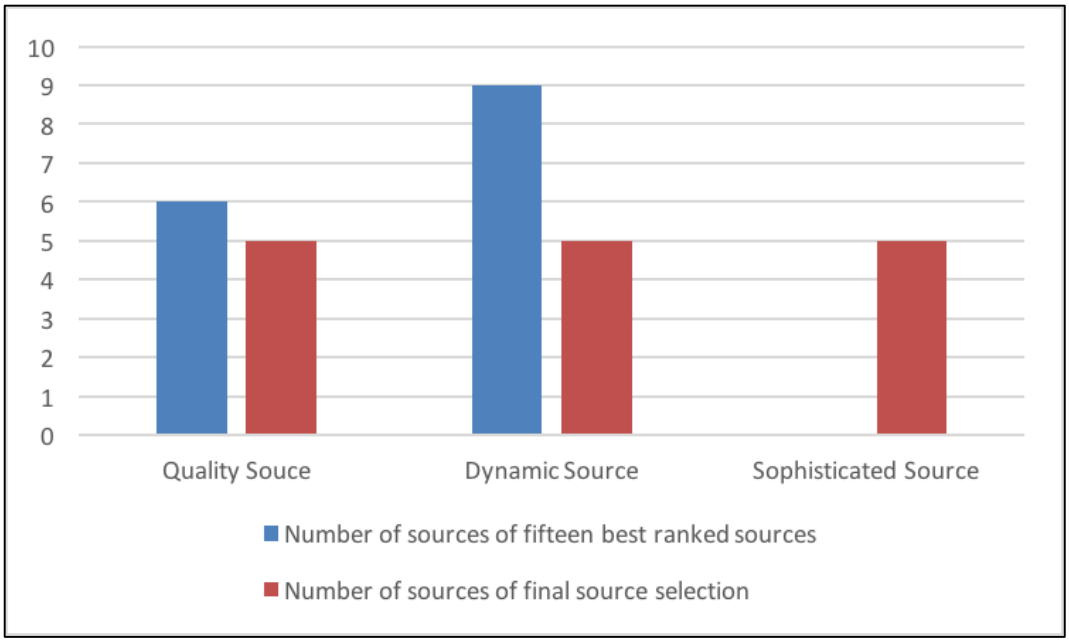


Figure 5.7: Data source category equilibrium performed on the single source evaluation results

	Category	*Source*	*Points in the value benefit analysis*
1	News & Techposts	Angel.co	336
2	News & Techposts	Crunchbase	336
3	News & Techposts	Silicon Republic	336
4	Scientific Publications	Science Direct	336
5	Scientific Publications	Scopus	336
6	Scientific Publications	Web of Science	336
7	Patents/Trademarks	Espacenet	331
8	News & Techposts	Kickstarter	329
9	Patents/Trademarks	EPO	326
10	News & Techposts	Interesting Engineering	321
11	Acc/Inc	Science Park Graz	295
12	Acc/Inc	Indibio	293
13	Acc/Inc	MainIncubator	288
14	Acc/Inc	Inbia	286
15	Festivals & Fairs	TED/TEDx	286

Table 5.5: Final selection of 15 high potential data sources

5.4 Data Processing

In this chapter, the sub-processes of the screen routine are described in detail. An overview of the process has been shown in chapter 5.1.2. The process consists of the following steps:

1. Initial keyword generation: Screening is similar to searching. Especially when the data set to be searched exceeds a critical amount of data, it becomes very important to know what one is searching for. It is unrewarding to search aimlessly. To overcome this, an initial keyword set is generated to start the search routine.
2. Data gathering from the selected sources: Since it is not possible to download all the data from the selected sources, the choice needs to be well defined. According to the initial keyword set, all relevant articles from the selected data sources are downloaded.
3. Natural Language Processing (NLP) of gathered data: After defining the relevant information, it needs to be analyzed. The aim of NLP is to filter and derive new keywords and concepts from the gathered articles. It smartly compresses the enormous amount of information downloaded into an amount of data that is easy to handle manually.
4. Steering the screen routine: Steering the routine means that, with the concepts and new keywords derived from the first search round, the initial keyword set can be adapted according to individual needs. Two directions are possible, focusing on a certain topic or a variety of topics.

A detailed illustration of the screen routine can be seen in Figure 5.8, where step five is not included, because it was described in chapter 5.1.2.

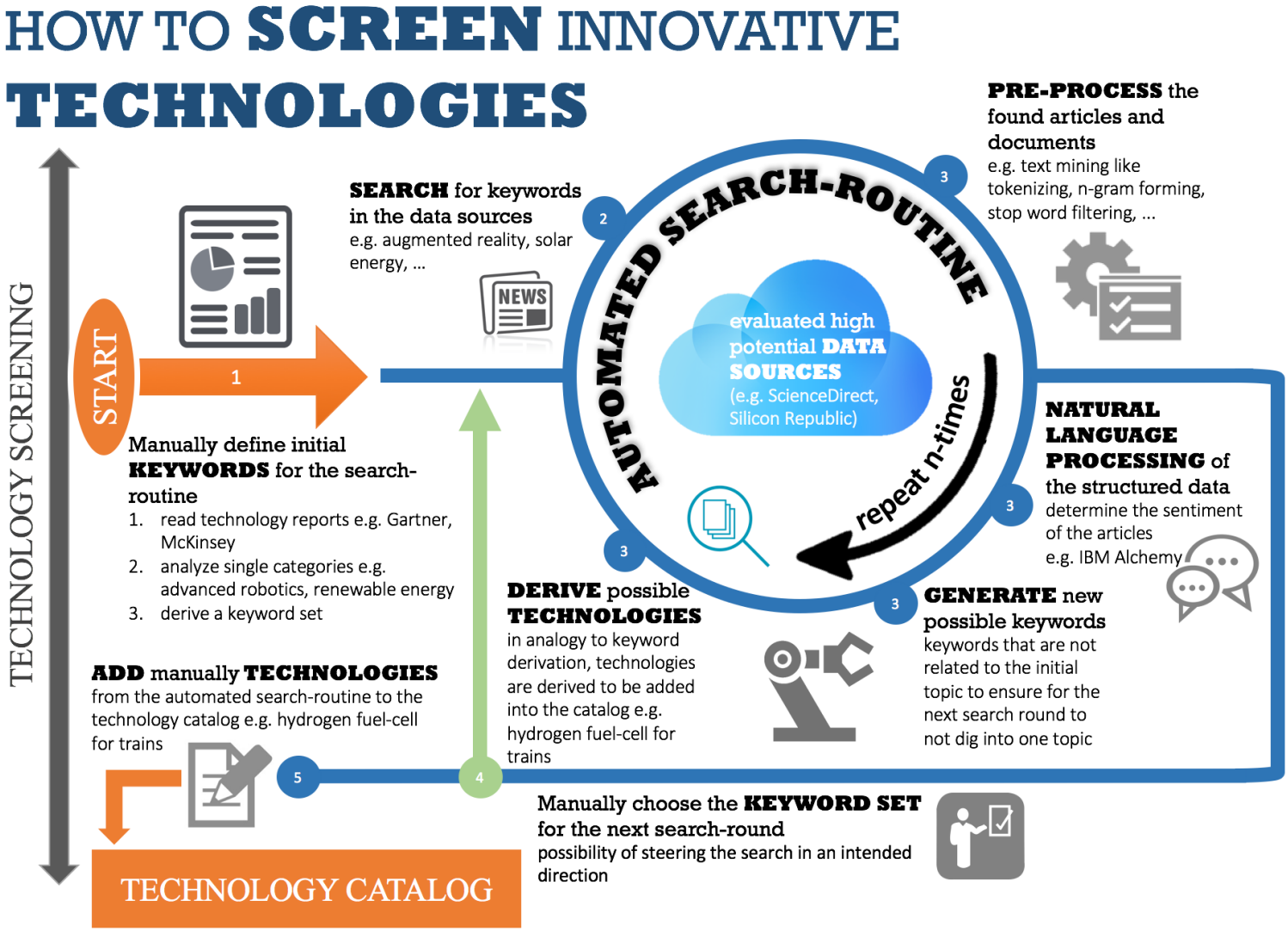


Figure 5.8: Flowchart of the developed screen routine approach

5.4.1 Initial Keyword Generation

The initial keyword set defines in which direction the first search round should go. Because the routine itself generates new keywords with every search round, the initial keyword set only needs to be defined for the first search. There are many different possibilities of how to come up with an appropriate keyword set. With this search routine, a keyword definition routine has been developed. The keywords generated by this routine are suitable for starting a general search on new technology trends. Depending on the topic the user searches for, the keyword set can also be defined by means of other processes, such as:

1. Individually define initial keyword set: For a manual definition, the user needs very detailed information on the task to be screened. In other words, the user searches for a solution to s/his problem and for this, s/he manually creates a list of keywords that best fit the problem statement.
2. Keyword brainstorm: If the problem statement is not as detailed, brainstorms are often used to define the initial set of keywords. The brainstorming group can consist of employees from different working areas. But for the brainstorming, the problem statement needs to be defined in some way as well.

The main focus of the process is the definition of an initial keyword set. As already mentioned, this process is perfectly viable if no search direction is given at all. Most of the time, if technology screening is conducted on a strategic level, input for the routine is quite rare, because nobody knows which potential substitutional technology a company's products and processes might influence. Thus, a very broad search needs to be established, focusing on future technology trends.

For this reason, looking back on the source evaluation chapter, there is one kind of source stated in the list that has not been used until now, because the source is too inert and difficult to analyze for the routine, technology reports from different consulting companies like McKinsey[246], Gartner[247], Boston Consulting[248] or Deloitte[249]. These reports are technology forecasts on a mid- and long term basis. The above mentioned consulting firms summarize high potential future technologies and their economic impact in those reports.

[246] http://www.mckinsey.com/
[247] http://www.gartner.com/technology/
[248] http://www.bcg.com/
[249] https://www2.deloitte.com/global/

The problem of these reports is that they are only published at long intervals and are therefore not updated as fast as needed for the screen routine. Additionally, they cannot be downloaded automatically, they need to be searched for manually and subsequently downloaded manually. These two reasons made them useless for the screen routine as data source.

Nevertheless, the information offered by these technology reports is of highest quality. They are multiple expert reviewed, contain ample information on the very topic interesting for identifying technology trends and the information is future-oriented. Meaning that, technology reports do not reflect the status quo of technology in use, but provide a forecast of relevant future technologies. Therefore, these technology reports build the basement for the initial keyword set generation.

In the example of the feasibility evaluation, one technology report is used: The Disruptive Dozen by McKinsey[250]. But for a broader perspective, and not being dependent on one single source, other technology reports can also be used to define the first keyword set.

In the case of the Disruptive Dozen[251], there are twelve chapters on the following different topics:

1. Mobile Internet
2. Automation of Knowledge Work
3. Internet of Things
4. Cloud Technology
5. Advanced Robotics
6. Autonomous Vehicles
7. Next Generation Genomics
8. Energy Storage
9. 3D Printing
10. Advanced Material
11. Advanced Exploitation of Oil and Gas Resources
12. Renewable Energy

[250] Manyika, et al. (2013)
[251] Manyika, et al. (2013)

Every chapter of the report contains approximately ten pages text. In the next step, these ten page sections are analyzed chapter by chapter by the NLP algorithm. For example, IBM Alchemy[252] can be used. After processing one chapter, the result looks as depicted in Figure 5.9. The left column (target) represents the derived concepts of the text. The middle column (type) shows the polarity of this concept (either positive or negative, as described in chapter 4.1.3.3). The right column (sentiment) gives the calculated relation to the text's sentiment.

Target	Type	Sentiment
mobile internet	positive	0.24818
mobile internet technology	positive	0.287022
mobile Internet access	positive	0.135999
mobile Internet usage	positive	0.213612
mobile devices	positive	0.0209316
mobile Internet devices	positive	0.173506
potential economic impact	positive	0.00922796
percent	negative	-0.108225
mobile computing devices	positive	0.435846
mobile Internet services	positive	0.476804
mobile Internet technologies	positive	0.412153
mobile Internet capability	positive	0.379027
potential mobile Internet	neutral	
new Internet users	positive	0.293534
mobile Internet applications	neutral	
new mobile services	positive	0.390641
New mobile software	positive	0.726044
wearable devices	positive	0.439985

Figure 5.9: Keywords generated with IBM Alchemy[253] on the chapter Mobile Internet of the McKinsey Disruptive Dozen[254]

[252] https://www.ibm.com/watson/
[253] https://www.ibm.com/watson/
[254] Manyika, et al. (2013)

From this list of concepts, the best suitable now need to be chosen manually. According to this selection, the first search will be started. In the above example, the following keywords were chosen for the first search round:

1. mobile internet access
2. mobile devices
3. mobile internet services
4. wearable devices

More words could be added to the search, or single words could be adapted. The goal of analyzing the technology reports is to provide the screen routine with first search direction. Therefore, the user is not obliged to use the terms suggested by the NLP algorithm.

This process is repeated for each chapter and the initial keywords are collected in a keyword set. For the feasibility analysis, the first initial keyword set consisted of 34 keywords and can be seen in Table 5.6.

Category	***Keyword***
3D Printing	additive manufacturing
Advanced Robotics	advanced robotics
Autonomous Vehicles	autonomous driving
Autonomous Vehicles	autonomous guidance
Autonomous Vehicles	autonomous systems
Cloud Technologies	cloud computing
Cloud Technologies	cloud services
Cloud Technologies	cloud technology
3D Printing	direct manufacturing
Energy Storage	electric vehicles
Energy Storage	energy storage technology
Next Generation Genomics	next-generation gene sequencing
Next Generation Genomics	next-generation genomics technology
Advanced Materials	graphene
Energy Storage	grid storage
Internet of Things	internet connected sensors
Automation of Knowledge Work	knowledge automation tools
Automation of Knowledge Work	knowledge work automation

Automation of Knowledge Work	knowledge worker occupations
Mobile Internet	mobile computing devices
Mobile Internet	mobile internet
Advanced Materials	carbon nanotubes
Renewable Energy	offshore wind
Advanced Materials	quantum dots
Internet of Things	real-time patient data
Renewable Energy	renewable energy
Internet of Things	rfid tags
Advanced Robotics	domestic service robots
Advanced Oil and Gas Exploration	shale gas reserves
Advanced Robotics	ultraprecise surgical robots
Next Generation Genomics	synthetic biology
Advanced Oil and Gas Exploration	unconventional oil
Advanced Oil and Gas Exploration	unconventional reserves
Mobile Internet	wearable devices

Table 5.6: Keyword set generated from the McKinsey Disruptive Dozen[255] report for the routines feasibility

Having defined the keywords to set off the screen routine, the next step is to gather data from the chosen sources.

5.4.2 Data Gathering within the Selected Sources

After having defined the target, the automated data processing can start. The data gathering stands at the beginning of this process. In chapter 5.3, we evaluated different sources and compiled them into a database set that provides us with diverse information with the necessary properties. In the previous chapter, we defined keywords to search for among these sources.

Now, the task is to download all the information relevant according to the defined keywords that can be found in the given sources.

Newsletter form one option of information downloading. Relevant news are frequently sent to the user via email and the data is gathered automatically by the users' subscription to the newsletter.

[255] Manyika, et al. (2013)

Query websites are another possibility. Query websites are common, especially when it comes to structured data like weather forecasts (temperature, humidity, rain, …) or stock prices. For unstructured data like news, articles and textual data, query websites are less common, but can still be found. The advantage of query websites is that the information gathering is nearly as simple as with newsletter subscription. Query websites have a predefined format, with which custom-written programs can query the database for information changes .

For more advanced web content downloading, web crawling tools were developed. These algorithms are fed with certain boundary conditions, for example, search words, depth of search (how many pages in a row can be opened without returning to the starting point). According to these boundary conditions, the algorithm searches the defined websites and downloads all data fitting the requirements. There are two major drawbacks. One is that many websites detect such crawling algorithms and block the access of these. The second one is that crawling is a more or less guided search. Great amounts of content are downloaded and therefore the search takes more time.

The last, most sophisticated and at the same time most work intensive approach is a customized downloader. A customized downloader is written to fit only the page the user wants to search and download. This downloader is used in the test routine to show the feasibility. For this reason, a customized downloader for the scientific publication website Science Direct[256] is written. The code can be found in Appendix 10.3.

5.4.3 Natural Language Processing of Gathered Data

The NLP algorithm is a key factor for the feasibility of this routine and it has a major impact on the performance of the routine.

For this thesis, several possibilities for distilling the data amount are tested. Most of them rest on the principles of text mining. Three different platforms are tested:

I: Rapid Miner[257]: Is a simple toolbox of text mining tools. These tools can be individually arranged depending on the needs of the operator. The toolset is extendible and offers many possibilities. There are three main reasons why Rapid Miner[258] was not chosen to be used for the screen routine:

[256] http://sciencepark.at/
[257] https://rapidminer.com/
[258] https://rapidminer.com/

1. Rapid Miner[259] is not able to handle big amounts of data. As soon as thousands of abstracts are processed by Rapid Miner[260], the program gets stuck and no results are delivered.
2. The output possibilities are limited.
3. Rapid Miner[261] lacks integration into other programming platforms. Since the aim is a basement for further developing the screen routine, Rapid Miner[262] is no viable base for further research, as it offers no possibility of integration into other programming platforms.

II: Matlab[263]: With this tool, the last drawback of Rapid Miner[264], lack of integration, is solved. The handling of big data amounts works well also with Matlab[265]. Only the performance of the data distillation was not satisfactory with Matlab[266]. At this point it became clear, that basic text mining tools do not deliver satisfactory results for this purpose.

It is not sufficient to compare documents, tokens and meanings. The data analyzing algorithm needs to focus more narrowly. It needs to detect a text's sentiment, because, by getting the sentiment, the algorithm is capable of not just filtering words, but finding the concept behind the abstract.

III: NLP algorithms: These algorithms get the sentiment of a text and derive concepts, entities and keywords from this sentiment. The algorithms are not limited to the words used within the text being processed. But further details on Natural Language Processing can be found in chapter 4.2

The chosen natural language processing tool for this thesis is IBM Bluemix[267]. Bluemix[268] offers a text processing library called IBM Alchemy[269]. This library rests on the basis of the IBM Watson[270]. IBM Alchemy[271] proved to be the right choice, which

[259] https://rapidminer.com/
[260] https://rapidminer.com/
[261] https://rapidminer.com/
[262] https://rapidminer.com/
[263] http://de.mathworks.com/
[264] https://rapidminer.com/
[265] http://de.mathworks.com/
[266] http://de.mathworks.com/
[267] https://www.ibm.com/watson/
[268] https://www.ibm.com/watson/
[269] https://www.ibm.com/watson/
[270] https://www.ibm.com/watson/
[271] https://www.ibm.com/watson/

can be read in chapter 5.6.2, where a comparison is made between a human distilling data and IBM Alchemy[272] doing the same task.

Another advantage of NLP is that it combines the following steps shown in figure Figure 5.8 data pre-processing, NLP, generation of new possible keywords and derivation of new technologies.

The result of the NLP is a list of concepts, entities and keywords that can be reviewed manually. A ten-page excerpt of this result can be found in Appendix 10.4.

5.4.4 Steering the Screen Routine

The last two steps, data gathering and the NLP of the gathered data, deal with huge quantities of data, which need to be fully automated. Through the NLP the data was compressed to a manually reviewable amount and hence steering the routine is done manually.

As already mentioned in the previous chapter the result of the data compression is a list of entities, concepts and keywords derived from the textual data fed into the routine. This list now needs to be reviewed manually. By doing so, the operator needs to watch out on two things:

1. New keywords for the next search round to steer the routine into the intended direction.
2. Technologies, trends, concepts being interesting for his employer to, enter those into the technology catalog.

Generally, it is recommended to start with spotting interesting technologies, trends and concepts, because during this process, the operator gets a better feeling in which direction the search went. For example if s/he spots an interesting technology in the list, s/he puts it into the technology catalog. This catalog is the final result of the screen routine, explained in chapter 5.5. Important for the operator, everything that seems relevant for his/her field of operation must be copied into the technology catalog. Hence this catalog is growing every search round.

After spotting relevant data from the compressed data list, the operator got a feeling in which direction the first search round went. By steering the routine the user can decide to further dig deeper into a topic, or pic a newly discovered technology and steer the search in the direction of this. By choosing the keyword set for the next search round the operator sets this direction.

---

[272] https://www.ibm.com/watson/

Detailed examples on steering the routine can be seen in chapter 5.6.3. There are keywords stated for both digging deeper into a topic on the one hand and wandering to new technologies on the other hand.

5.5 Technology Catalogue

As already mentioned the technology catalog is the result of the screen routine. It sums up all the relevant technologies and trends discovered during the screening process. Every search round the interesting concepts are entered into the catalog manually.

This catalog sets the basement for further processing or simply as information communication tool. The technology catalog can be presented to strategic management employees so they can make their calls for action according to this found information. Or it can also be the basement for further processing. This can be assessment processes for R&D departments, competitor analysis or acquisition decisions.

An example for a technology catalog can be seen in Table 5.7. There are ten technologies entered. The whole technology catalog generated by the screen routine during it’s feasibility test can be seen in the Appendix 10.5.

Category	*Technology*	*Application*	*Source*	*Year*	*URL/Abstract*
Construction	3d printing	on site 3d-printing of houses	Interesting Engineering	2016	http://interestingengineering.com/3d-printed-office-is-the-office-of-the-future/
Advanced Production	3d printing	carbon-fiber compounds	Science Direct	2015	Investigation into the Development of an Additive Manufacturing Technique for the Production of Fibre Composite Products
Advanced Production	3d printing	additive manufacturing for shape memory polymer	Science Direct	2015	Characterization of polyurethane shape memory polymer processed by material extrusion additive manufacturing
Advanced Production	3d printing	solid freeform fabrication	Science Direct	2016	The cost of additive manufacturing: machine productivity, economies of scale and technology-push
Robotics	acoustic source localization	robot orientation	Science Direct	2003	AR_service-robotics_Abstr98
Health	alginate quantum dots	gene delivery	Science Direct	2016	Cationic carbon quantum dots derived from alginate for gene delivery: One-step synthesis and cellular uptake
Computing	ambient intelligence	grid-computing	Science Direct	2014	The Internet of Things vision: Key features, applications and open issues
Mobility	antimatter propulsion	space travel	Kickstarter	2016	https://www.kickstarter.com/projects/2114765394/antimatter-propulsion?ref=category_popular
Computing	artificial intelligence	machine encryption	Techcrunch	2016	https://techcrunch.com/2016/10/28/googles-ai-creates-its-own-inhuman-encryption/
Health	artificial intelligence	intelligent health diagnostics	Silicon Republic	2016	https://www.siliconrepublic.com/start-ups/kinesis-medtech-funding

Table 5.7: Example for a technology catalog with 10 entries

The following columns are part of the technology catalog:

1. Category: Every technology is classed to a category, in order to ease later on evaluation or filtering processes.
2. Technology: In the technology column the relevant technology or trend found in the generated list is stated there.
3. Application: For a technology in general more than one application exists and also the other way around. For example the application of propelling a car, can be either solved by the technology of internal combustion engines or by electric motors. The technology of an electric motor as well as the internal combustion engine on the other hand can also be used for other applications, like propelling motorbikes or other things.
4. Source: This states the source, where the information comes from, this gets interesting as soon as the quality of the source is involved. Because technologies of less reliable sources might need a backup check on the correctness of information.
5. Publication Year: There the year, when the article, from where the information was retrieved is stated. This is also interesting for further processing.
6. URL: There the URL to the article is stated where the information was retrieved. This is important if anyone wants to inform his/herself on the technology.

5.6 Demonstration of Properties

According to the desired properties (shown in chapter 5.2), in this chapter the demonstration in form of examples can be seen. The examples are from the feasibility of the developed screen routine.

5.6.1 Data Gathering

As already mentioned in chapter 5.2.1 for the demonstration of properties and the routines feasibility only one data source is used. The programming effort to write a customized downloader is high and needs to be repeated for every new source being implemented.

The chosen source for the routines feasibility is Science Direct[273]. There global scientific publications are stored in a database. Nether the less a second source has been tried to be implemented. The second source are the final thesis of the TU Graz.

[273] http://www.sciencedirect.com/

The reason why this second source is not part of the routines feasibility is because of the too small amount of data available. For many keyword searches the final thesis database of the TU Graz did not deliver enough documents for further processing. This customized downloader for the TU Graz is coded for the field experiment explained in chapter 5.1.1, where similar results can be observed. Therefore the data gathering concentrates on Science Direct[274] only.

Due to the license's limitation of the TU Graz, the customized algorithm can only download the first 1000 documents of a keyword search. The code was written in Matlab R2015a[275] and can be found in Appendix 10.3.

Nevertheless, the principle of the code is described here. The customized web downloader for Science Direct[276] does the same thing a human would do. The website link (URL) of the keyword search needs to be entered in the Matlab[277] script (the URL of the first two pages of the search result), this can be seen in Figure 5.10. The script opens the URL and retrieves every abstract of the website and saves it as a text file. After downloading all abstracts of one page, it proceeds to the next one.

```
%- %-http://www.sciencedirect.com/science?
_ob=ArticleListURL&_method=list&_ArticleListID=-1095774972&_sort=r&_st=13&view=c&md5=
a18d807a836183a0dcf153b6bf86ca5d&searchtype=a-% -%

%- %-http://www.sciencedirect.com/science?
_ob=ArticleListURL&_method=tag&searchtype=a&refSource=search&pdfDownloadSort=r&PDF_DD
M_MAX=25&_st=13&count=1000&sort=r&filterType=&_chunk=0&hitCount=7715&NEXT_LIST=1&view
=c&md5=65e10b2dd6022b50ba7014b016c6d865&_ArticleListID=-1095774972&chunkSize=25&sisr_
search=&TOTAL_PAGES=309&pdfDownload=&zone=exportDropDown&citation-
type=RIS&format=cite-abs&bottomPaginationBoxChanged=&bottomNext=Next+%3E
%3E&displayPerPageFlag=f&resultsPerPage=25-% -%
```

Figure 5.10: Input text file for customized Science Direct[278] downloader with the URLs of the first two search result pages for Matlab[279]

Unfortunately, due to an unpredictible anomaly on the Science Direct[280] website, the data gathering stopped at random points in some keyword searches. As already mentioned, due to licensing, not more than 1000 abstracts per keyword search are possible, but because of this anomaly, some keyword searches already stopped downloading after a few hundred abstracts. This anomaly could not be solved before

[274] http://www.sciencedirect.com/
[275] http://de.mathworks.com/
[276] http://www.sciencedirect.com/
[277] http://de.mathworks.com/
[278] http://www.sciencedirect.com/
[279] http://de.mathworks.com/
[280] http://www.sciencedirect.com/

this thesis was concluded. Nevertheless, the data amount gathered from Science Direct[281] was sufficient.

According to the initial keyword set, which was mined by the McKinsey Desruptive Dozen[282], the following keyword searches with the number of downloaded abstracts can be seen in Table 5.8.

Category	*Keyword*	*Science Direct search word*	*Number of downloaded abstracts*
3D Printing	additive manufacturing	additive manufacturing	**425**
Advanced Robotics	advanced robotics	advanced robotics	**125**
Autonomous Vehicles	autonomous driving	autonomous driving	**150**
Autonomous Vehicles	autonomous guidance	autonomous guidance	**600**
Autonomous Vehicles	autonomous systems	autonomous system	**1000**
Cloud Technologies	cloud computing	cloud computing	**1000**
Cloud Technologies	cloud services	cloud services	**425**
Cloud Technologies	cloud technology	cloud technology	**1000**
3D Printing	direct manufacturing	direct manufacturing	**475**
Energy Storage	electric vehicles	electric vehicles	**475**
Energy Storage	energy storage technology	energy storage technology	**850**
Next Generation Genomics	next-generation gene sequencing	gene sequencing	**325**
Next Generation Genomics	next-generation genomics technology	genomics	**225**
Advanced Materials	graphene	graphene	**1000**
Energy Storage	grid storage	grid storage	**850**
Internet of Things	internet connected sensors	internet connected sensors	**275**
Automation of Knowledge Work	knowledge automation tools	knowledge automation tool	**1000**
Automation of Knowledge Work	knowledge work automation	knowledge work automation	**650**
Automation of Knowledge Work	knowledge worker occupations	knowledge worker occupation	**450**
Mobile Internet	mobile computing devices	mobile computing device	**800**
Mobile Internet	mobile internet	mobile internet	**1000**

[281] http://www.sciencedirect.com/
[282] Manyika, et al. (2013)

Advanced Materials	carbon nanotubes	nanotubes	**1000**
Renewable Energy	offshore wind	offshore energy	**1000**
Advanced Materials	quantum dots	quantum dots	**1000**
Internet of Things	real-time patient data	Real-time patient data	**550**
Renewable Energy	renewable energy	renewable energy	**1000**
Internet of Things	rfid tags	RFID Tags	**400**
Advanced Robotics	domestic service robots	service robot	**100**
Advanced Oil and Gas Exploration	shale gas reserves	shale gas	**450**
Advanced Robotics	ultraprecise surgical robots	surgical robot	**100**
Next Generation Genomics	synthetic biology	synthetic biology	**550**
Advanced Oil and Gas Exploration	unconventional oil	unconventional oil	**175**
Advanced Oil and Gas Exploration	unconventional reserves	unconventional reserves	**225**
Mobile Internet	wearable devices	wearable device	**250**

Table 5.8: Number of downloaded abstracts referred to the defined search words

The search resulted in a total of 19825 of downloaded abstracts. The big challenge is to compress this data amount into manually reviewable amounts of high quality. Even if we defined the Big Data analysis to be kept in the range of terabytes, which the 19825 abstracts do not reach, after implementing enough sources, it is only a matter of time before this routine leapt into the field of Big Data analytics. Nevertheless, for the feasibility, this is not necessary to show, but with the Science Direct[283] customized downloader the feasibility of collecting relevant data can be shown.

5.6.2 Data Distillation

Data distillation is a key success factor for the screen routine, therefore data distillation was extensively tested. Three different methods for compressing data are tested. These three approaches are described in chapter 5.4.3. For demonstration purposes, the following example shows how good IBM Alchemy[284] adapts to the needs of this thesis.

[283] http://www.sciencedirect.com/
[284] https://www.ibm.com/watson/

For this example, the first chapter of the McKinsey technology report 'The Disruptive Dozen'[285] was mined, once manually and once with IBM Alchemy[286]. The first chapter, Mobile Internet, consists of roughly 4700 words.

In the first step, the text was read through and all keywords, concepts, entities, that can be interesting for the further routines process were marked manually. This manual process can be seen in Figure 5.11. All terms marked in yellow are concepts or keywords for the next search round, or entered directly in the technology catalog. In other words, all terms that should not get lost by data compression are marked in yellow in Figure 5.11.

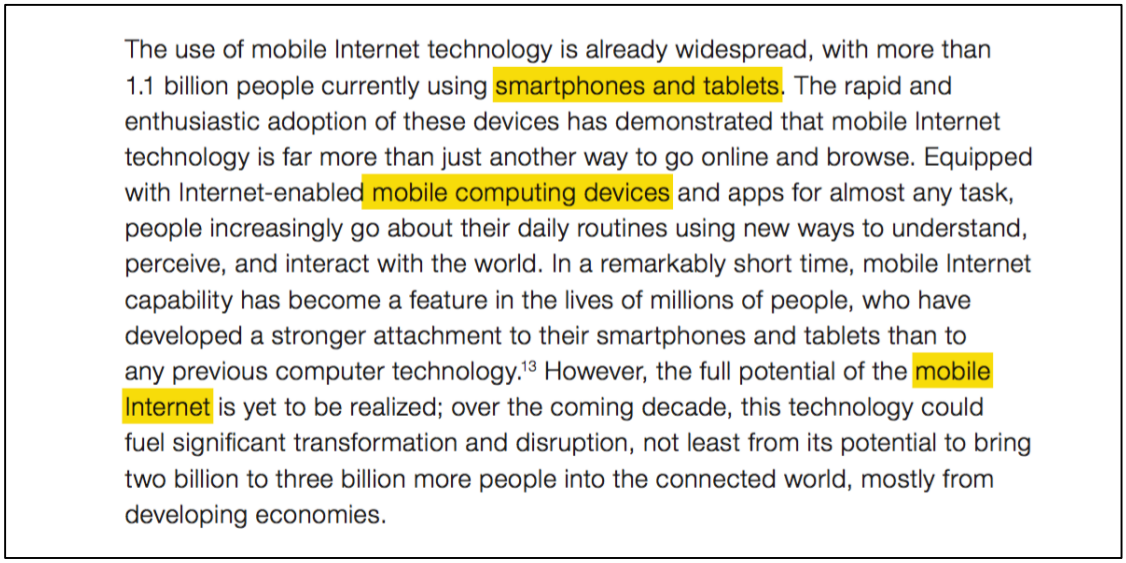

The use of mobile Internet technology is already widespread, with more than 1.1 billion people currently using smartphones and tablets. The rapid and enthusiastic adoption of these devices has demonstrated that mobile Internet technology is far more than just another way to go online and browse. Equipped with Internet-enabled mobile computing devices and apps for almost any task, people increasingly go about their daily routines using new ways to understand, perceive, and interact with the world. In a remarkably short time, mobile Internet capability has become a feature in the lives of millions of people, who have developed a stronger attachment to their smartphones and tablets than to any previous computer technology.[13] However, the full potential of the mobile Internet is yet to be realized; over the coming decade, this technology could fuel significant transformation and disruption, not least from its potential to bring two billion to three billion more people into the connected world, mostly from developing economies.

Figure 5.11: Manually defining keywords in the first chapter Mobile Internet of the McKinsey Disruptive Dozen[287] (an excerpt)

The result of this manual data compression can be seen in the following list; these are the terms that should not get lost during the data distillation according to the manual review:

- Smartphones and tablets
- Mobile computing devices
- Mobile internet
- Wearable devices
- 4G wireless networks
- Health monitoring

[285] Manyika, et al. (2013)
[286] https://www.ibm.com/watson/
[287] Manyika, et al. (2013)

- Cloud access
- High speed wireless connectivity
- Smart watches
- Cellular networks
- Long-range Wifi
- Dynamically sharing spectrum
- Internet based services
- Near-field payments
- Augmented reality
- Personal assistant
- Virtual reality glasses
- Lithium batteries

This list is compared to the keywords generated by IBM Alchemy[288]. The list of keywords with the corresponding polarity and intensity estimations generated by IBM Alchemy[289] can be found in Figure 5.9.

By comparing both lists, the following terms were correctly identified as keywords by IBM Alchemy[290]:

- Mobile computing devices
- Mobile Internet
- Wearable Devices
- Internet Based Services

In other words, these four keywords were spotted both with manual and algorithm-based processing. These are the targeted hits. IBM Alchemy[291] generated 18 keywords. This is less than 4 ‰ of the 4700 words in total. From this 4 ‰, more than 20 % of the 18 keywords are similar to the ones chosen manually.

The results of this comparison are summed up and illustrated in Figure 5.12. As can be seen, four terms are coincidences between IBM Alchemy[292] and the manual

[288] https://www.ibm.com/watson/
[289] https://www.ibm.com/watson/
[290] https://www.ibm.com/watson/
[291] https://www.ibm.com/watson/
[292] https://www.ibm.com/watson/

keyword generation. Nevertheless, IBM Alchemy[293] found six keywords that are suitable for the routine's further process, but which were not spotted manually.

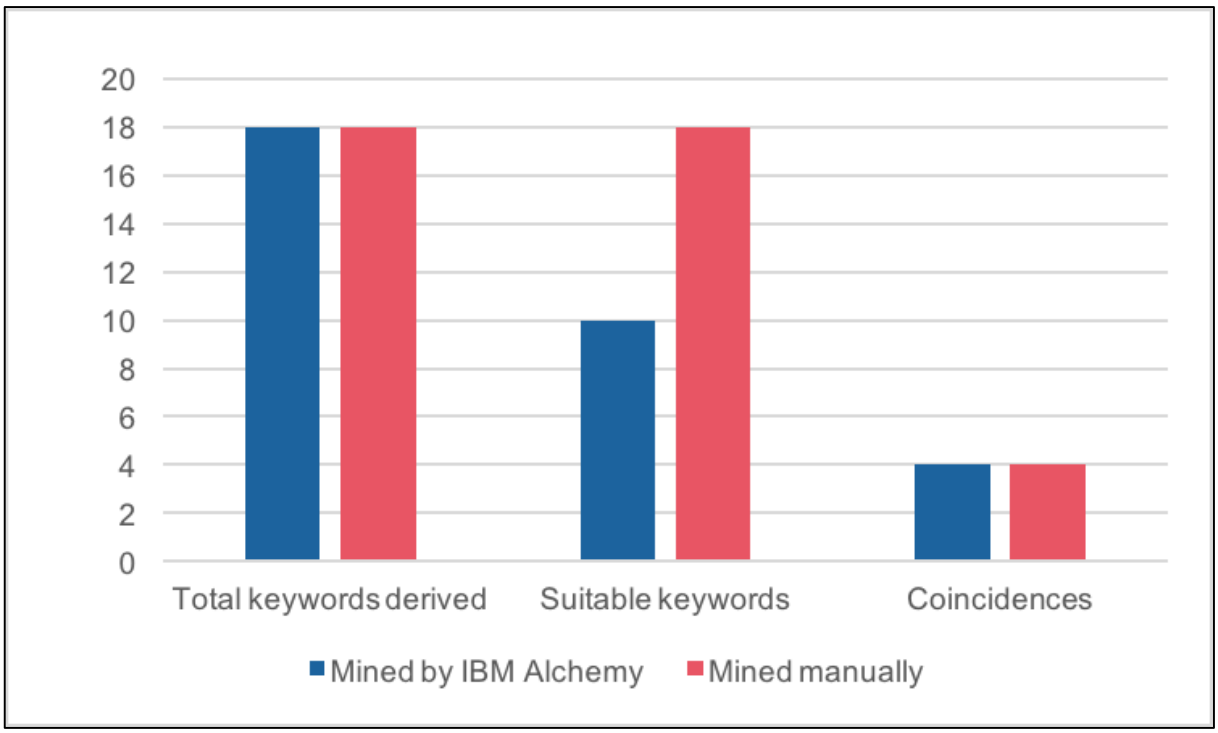


Figure 5.12: IBM Alchemy[294] vs. manual Search (on the example of McKinsey Disruptive Dozen[295] chapter Mobile Internet)

There is also a second example worth mentioning. IBM Alchemy[296] does not simply filter the words occurring within the text being processed, it derives the sentiment of a text and, according to this, generates concepts that fit this sentiment. Due to this it is possible that keywords generated via IBM Alchemy[297] do not occur in the text. This behavior can be best seen in the following example. The results for IBM Alchemy[298] mining of this text can be seen in Figure 5.13.[299]

[293] https://www.ibm.com/watson/
[294] https://www.ibm.com/watson/
[295] Manyika, et al. (2013)
[296] https://www.ibm.com/watson/
[297] https://www.ibm.com/watson/
[298] https://www.ibm.com/watson/
[299] https://www.ibm.com/watson/

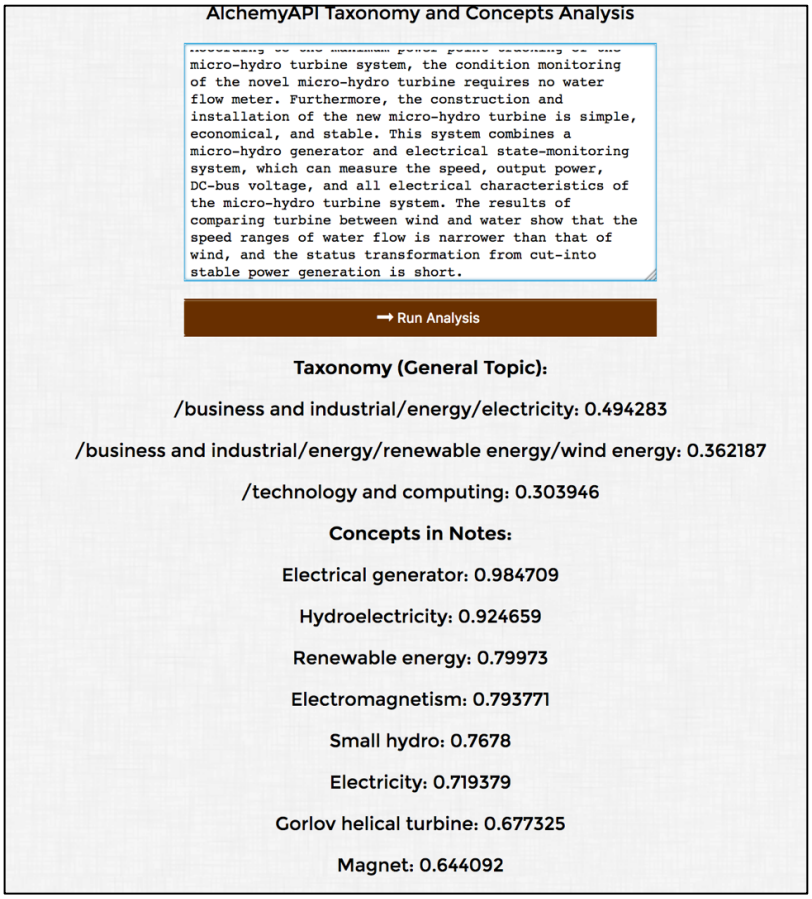


Figure 5.13: Keyword generation of IBM Alchemy[300] on an abstract on three-bladed vertical hydro turbines

The second last concept is interesting, which is the Gorlov helical turbine. In the text, the word Gorlov does not occur, nevertheless, IBM Alchemy[301] detected the sentiment of the text and computed the concepts behind that. It made the correct interpretation that this text is about a three-bladed Gorlov turbine.

These two examples show why IBM Alchemy[302] best fits the requirements of the developed screen routine.

The implementation of IBM Alchemy[303] into the routine's feasibility was done via PythonIII[304] in combination with IBM Bluemix[305]. The Python[306] script can be seen in Appendix 10.6.

---

[300] https://www.ibm.com/watson/
[301] https://www.ibm.com/watson/
[302] https://www.ibm.com/watson/
[303] https://www.ibm.com/watson/
[304] https://www.python.org/
[305] https://www.ibm.com/watson/
[306] https://www.python.org/

5.6.3 Routine Maneuverability

If the operator chooses certain keywords from the generated list, the routine can be steered in the intended direction. This selection of keywords for the different search rounds can be seen as a keyword strategy.

An example can be seen in Table 5.9. All keywords listed in this table are generated by the routine and stated in the list of compressed information. The first column (of the example table) shows the initial topics the routine searched in. The second column gives a keyword to focus more narrowly on the initial topic and the right column gives a keyword for examining a variety of topics. For example, taking the initial topic of advanced material, a keyword further examining this topic can be carbon nanotubes, suggested by the screen routine. By choosing the keyword solar cells, the screen routine will wander in the direction of renewable energy. In this example, only single keywords were chosen; however, for steering the routine multiple keywords must be changed towards the intended direction.

Initial topic	*Keyword to dig deeper*	*Keyword to find new topic*
renewable energy	biofuels	Haber-Bosch process
advanced material	carbon nanotubes	solar cells
alternative oil and gas exploration	shale gas	artificial intelligence
3d-printing	equal channel angular extrusion	cybernatics

Table 5.9: Keyword strategy for the subsequent search round

By choosing the right keyword strategy, the routine will move in the intended direction in the next search round.

5.6.4 Easy and Intuitive Operation

To prove operation is easy and intuitive, the routine is not developed far enough. The basic concept of a semi-automated routine lays the perfect foundations for an intuitive operation. In addition, the routine's maneuverability, as already described in chapter 5.2.2, is developed for easy and intuitive operation. Nevertheless, the final and fully programmed routine can be proven to be operable intuitively.

5.6.5 Routine Speed

As is the case with easy and intuitive operation, the feasibility of the routine's speed cannot be proven until the routine is fully programmed. The data source evaluation shown in chapter 5.3 ensures minimization of the number of needed sources so one-sided information can be avoided and therefore is a good base for sufficient routine speed. Nevertheless, depending on the algorithms used to download and compress the data, the routine's speed can be improved dramatically.

6 Conclusion & Outlook

This thesis was initiated with the identification of the need for structured tools that can analyze considerable amounts of data. The characteristics of such a routine are derived from needs of potential users. Useful requirements for the screen routine are additionally collected by means of literature review and a field experiment.

This know-how forms the basis for the design of the screen routine approach. By implementing the required parts of this approach step by step, the feasibility of the key concepts of the routine are shown.

The routine development started with the identification of different sources. By reviewing literature data sources are gathered and evaluated with a value benefit analysis. Since the screen routine is fed with more than one source, the characteristics of all selected sources are crucial for the routine's performance. For this reason three categories of sources are identified.

For the information gathering process a customized downloader was programmed, that works for both, the TU Graz intranet and Science Direct[307]. With the help of this program, thousands of abstracts can be downloaded automatically.

For the data distillation IBM Alchemy[308] is used. Via a PythonIII[309] script the IBM[310] library was implemented and the gathered information is analyzed.

As a result of the screen routine a technology catalog is generated. There, interesting technologies are stated and summarized together with the source, publication year and the application of the technology. This technology catalog can be used as input to further analyzing tools.

By means of the different examples processed through the screen routine, the feasibility for the desired properties has been shown. The routine gathers and compresses data as intended and therefore enables an efficient steering mechanism. Through this steering mechanism the screen routine can be operated in different fields of application like venture capital market, merger and acquisition, research and development and strategic decision making.

Within this research the feasibility of the developed screen routine has been shown for the desired characteristics. Nether the less further research is needed to fully reveal

---

[307] http://www.sciencedirect.com/
[308] https://www.ibm.com/watson/
[309] https://www.python.org/
[310] https://www.ibm.com/watson/

the routine's feasibility, especially for the characteristics depending on the routine being programmed as a whole, such as:

1. Routine speed
2. Easy and intuitive operation

As a result, close cooperation with Big Data and text mining experts becomes necessary. With a fully programmed model of the routine, the process of showing the feasibility can be finished. Furthermore, this mockup can be used for the first beta-test users. The input of these early users will ensure an increase in the quality of the routine as well as fulfillment of the requirements of potential users.

Additionally, the list of sources must be further adapted. Since websites are steadily changing, this data source list also needs to be updated continuously. By adding new sources and refining the data selection process, the routine's performance can be further increased and custom-tailored to the requirements of single users.

Furthermore additional data sources need to be implemented into the routine. Similar to the Science Direct[311] downloader, other sources from different categories need to be implemented. By doing so, further research on data gathering is suggested, in best case with data analysis experts.

In parallel, a business plan needs to be conceived, to find out how best to exploit the potential of the findings in this thesis. Within this business plan a close market research needs to be conducted in order to clearly spot potential competitors.

[311] http://www.sciencedirect.com/

7 List of Figures

8 List of Tables

9 List of References

Aggarwal, C., & Zhai, C. (2012). *Mining Text Data.* New York: Springer.

Amezcua-Martínez, J., & Güemes-Castorena, D. (2010). *Strategic Foresight Methodology to Identifying Technology Trends and Business Opportunities.* Monterrey: IEEE.

Bachmann, R., Kemper, G., & Gerzer, T. (2014). *Big Data - Fluch oder Segen? Unternehmen im Spiegel gesellschaftlichen Wandels.* Frechen: MITP.

Banchs, R. E. (2013). *Text Mining with MATLAB.* New York, NY: Springer New York.

Bhanuse, S., Kamble, S., & Kakde, S. (2015). *Text Mining using Metadata for Generation of Side information.* Nagpur: Elsevier.

Boe-Lillegraven, S., & Monterde, S. (2014). *Exploring the cognitive value of technology foresight: The case of the Cisco Technology Radar.* Aarhus: Elsevier.

Burgelman, R. A., Christensen, C. M., & Wheelwright, S. C. (2005). *Strategic management of technology and innovation.* Boston, Mass: McGraw-Hill.

Carstensen, K.-U., Ebert, C., Ebert, C., Jekat, S., & Klabunde, R. (2010). *Computerlinguistik und Sprachtechnologie Eine Einführung.* Heidelberg: Spektrum.

Cetron, M. (1969). *Technological Forecasting: A Practical Approach.* London: Gordon and Breach.

Cetron, M. J. (1969). *Technological forecasting : a practical approach.* London: Gordon and Breach.

Dobrzañska-Danikiewicz, A. (2010). *E-foresight of materials surface engineering.* Gliwice: OCSCO World Press.

Ester, M., & Sander, J. (2000). *Knowledge Discovery in Databases Techniken und Anwendungen.* München: Springer.

Fasel, D., & Meier, A. (2016). *Big Data Grundlagen, Systeme und Nutzungspotenziale.* Zürich: Springer.

Hausser, R. (2000). *Grundlagen der Computerlinguistik Mensch -Maschine-Kommunikation in natürlicher Sprache.* Berlin: Springer.

Herbrich, R., & Graepel, T. (2010). *Handbook of Natural Language Processing.* Cambridge: Chapman & Hall/CRC.

Iafrate, F. (2015). *From Big Data to Smart Data.* London: ISTE.

Indurkhya, N., & Damerau, F. (2010). *Handbook of Natural Language Processing.* Boca Raton: Chapman & Hall.

Ittoo, A., Minh Nguyen, L., & van den Bosch, A. (2015). *Text analytics in industry: Challenges, desiderata and trends.* Belgien: Elsevier.

Kang, T.-H., Tsai, L.-M., & Horng, S. (2009). *A Study of Applying the Internet Platform on Technology Foresight.* Portland: PICMET.

Kochikar, V. (2008). *Addressing the 'Technology Foresight Defecit': A Multidimensional Approach.* IEEE.

Kochikar, V. P. (2008). *Addressing the 'Technology Foresight Deficit': A Multidimensional Approach.* IEEE.

Möhrle, M. (2008). *Technologie Roadmapping Zukunftsstrategien für Technologieunternehmen.* Berlin: Springer.

Manyika, J., Chui, M., Bughin, J., Dobbs, R., Bisson, P., & Marrs, A. (May 2013). Disruptive technologies: Advances that will transform life, business, and the global economy.

Marz, N., & Warren, J. (2015). *Big-Data Principles and best practices of scalable real-time data systems.* Shelter Island: Manning.

Nagl, R. (6. Dezember 2016). (P. Berner, Interviewer)

Ohlhorst, F. (2013). *Big Data Analytics: Turning Big Data into Big Money.* New Jersey: Wiley.

Pietrobelli, C., & Puppato, F. (2015). Technological Forecasting & Social Change. *Science Direct*.

R. Rohrbeck, J. H. (2006). *Technology Intelligence and Innovation Strategy.* Berlin: IEEE.

Ritchey, T. (2006). Problem Structuring using Computer-Aided Morphological Analysis. *Journal of the Operational Research Society, 57*(7).

Rohrbeck, R. (2011). *Coporate Foresight: Towards a Maturity Model for the Future Orientation of a Firm.* Berlin: Springer.

Schmarzo, B. (2013). *Big Data: Understanding How Data Powers Big Business.* Indianapolis: Wiley.

Schuh, G., & Klappert , S. (2011). *Technologie Management Handbuch Produktion und Management 2.* Berlin Heidelberg: Springer.

Tidd, J., & Bessant, J. (2013). *Managing Innovation: Integrating Technological, Market and Organizational Change.* Chichester: Wiley.

Verein Deutscher Ingenieure. (November 1998). Design engineering methodics - Engineering design at optimum cost- Valuation of costs. *Technische Regel*. VDI.

Vijay, K., & Balachandre, D. (2015). *Predictive Analytics and Data Mining.* Waltham, Mass.: Morgan Kaufmann.

10 Appendix

10.1 Possible Online Data Sources

Literature and Startup Review

Data Sources

Categorie	Source-Name	Rang
News & Techposts	News & Techposts | Angel.co	1
News & Techposts	News & Techposts | Crunchbase	2
News & Techposts	News & Techposts | Silicon Republic	2
Scientific Publications	Scientific Publications | Science Direct	2
Scientific Publications	Scientific Publications | Internet Scientific Publications	2
Scientific Publications	Scientific Publications | Int. Scientific Publications	2
Patents/Trademarks	Patents/Trademarks | Espacenet	7
News & Techposts	News & Techposts | Kickstarter	8
Patents/Trademarks	Patents/Trademarks | USPto	9
Patents/Trademarks	Patents/Trademarks | EPO	9
News & Techposts	News & Techposts | Techbrunch	11
News & Techposts	News & Techposts | Interesting Engineering	11
News & Techposts	News & Techposts | Wonderful Engineering	11
News & Techposts	News & Techposts | GeekWire	14
News & Techposts	News & Techposts | DailyTech	15

News & Techposts	News & Techposts | BBC	15
News & Techposts	News & Techposts | Wordpress	15
News & Techposts	News & Techposts | Science News	15
News & Techposts	News & Techposts | Der Brutkasten	19
News & Techposts	News & Techposts | Forbes	20
News & Techposts	News & Techposts | Der Standard	21
News & Techposts	News & Techposts | Die Presse	21
News & Techposts	News & Techposts | Süddeutsche	21
News & Techposts	News & Techposts | Gewinn	21
News & Techposts	News & Techposts | Wirtschaftsblatt	21
News & Techposts	News & Techposts | LifeScience	26
News & Techposts	News & Techposts | TechWorld	27
Acc/Inc	Acc/Inc | Science Park Graz	28
News & Techposts	News & Techposts | TechStars	28
Patents/Trademarks	Patents/Trademarks | TMDB	28
Acc/Inc	Acc/Inc | Indibio	31
News & Techposts	News & Techposts | Futurezone	31
Acc/Inc	Acc/Inc | MainIncubator	33
News & Techposts	News & Techposts | Trending-Topics	33
News & Techposts	News & Techposts | Die kleine Zeitung	33
Acc/Inc	Acc/Inc | Inbia	36

Universities	Universities | MU-Leoben	56
Festivals & Fairs	Festivals & Fairs | Lean Startup	59
News & Techposts	News & Techposts | Technologie.at	60
Acc/Inc	Acc/Inc | Oddup	61
Acc/Inc	Acc/Inc | LaunchBox	62

10.2 Value Benefit Analysis for Single Source Evaluation

Data Sources

Value Benefit Analysis

Source \ Criteria	Time-Lag	Frequency	Reginality	Amount of Relevant Data	Number of Authors	Reliability	Ease of Data Gathering	Points	Rank
Weight	4,5	5	1	4	2,5	3	1		
Relative weight	21,4	23,8	4,7	19,0	11,9	14,2	4,7		
Acc/Inc | Science Park Graz	3	3	1	3	3	4	1	295	29
Acc/Inc | LaunchBox	0	0	3	4	2	4	0	171	62
Acc/Inc | Inbia	3	2	4	3	3	4	1	286	37
Acc/Inc | MainIncubator	3	2	3	3	4	4	0	288	34
Acc/Inc | Indibio	3	2	3	3	4	4	1	293	32
Acc/Inc | Oddup	0	0	4	4	4	4	0	200	61
Festivals & Fairs | 15 Sec. Festival	2	1	3	3	3	4	1	236	51
Festivals & Fairs | Pioneers Festival	2	1	2	3	3	4	1	231	54
Festivals & Fairs | Ted/Tedx	2	3	4	2	4	4	2	286	37
Festivals & Fairs | Lean Startup	2	1	3	3	2	4	1	224	59
Festivals & Fairs | Startup Grind	2	2	4	3	4	4	1	276	44
Festivals & Fairs | EU-Startups	2	1	3	3	3	4	1	236	51
Festivals & Fairs | The Startup Conference	2	1	3	3	3	4	1	236	51
News & Techposts | Crunchbase	4	4	4	2	4	2	4	333	3
News & Techposts | Techbrunch	4	4	4	2	3	2	4	321	12
News & Techposts | Kickstarter	4	4	4	3	4	1	2	329	9
News & Techposts | Startablish	3	3	2	2	2	2	4	255	47
News & Techposts | TechStars	3	3	3	3	3	3	2	295	29
News & Techposts | Pioneers	2	1	2	3	3	4	1	231	54
News & Techposts | Der Brutkasten	4	4	2	2	3	2	4	312	20

News & Techposts | Technologie.at	2	2	1	2	2	3	2	210	60
News & Techposts | Interesting Engineering	4	4	4	2	3	2	4	321	12
News & Techposts | Wonderful Engineering	4	4	4	2	3	2	4	321	12
News & Techposts | Silicon Republic	4	4	4	2	4	2	4	333	3
News & Techposts | Trending-Topics	4	3	2	2	3	2	4	288	34
News & Techposts | GeekWire	4	4	3	2	3	2	4	317	15
News & Techposts | DailyTech	4	4	4	1	4	2	4	314	16
News & Techposts | BBC	4	4	4	1	4	2	4	314	16
News & Techposts | Futurezone	4	4	2	1	3	2	4	293	32
News & Techposts | Wordpress	4	4	4	1	4	2	4	314	16
News & Techposts | Angel.co	2	4	4	4	3	4	2	336	2
News & Techposts | TechWorld	4	4	4	1	3	2	3	298	28
News & Techposts | LifeScience	4	4	4	1	3	2	4	302	27
News & Techposts | Science News	4	4	4	1	4	2	4	314	16
News & Techposts | Ideentriebwerk Graz	4	3	1	2	2	2	2	262	45
News & Techposts | Der Standard	4	4	2	1	4	2	4	305	22
News & Techposts | Die Presse	4	4	2	1	4	2	4	305	22
News & Techposts | Die kleine Zeitung	4	4	1	1	3	2	4	288	34
News & Techposts | Süddeutsche	4	4	2	1	4	2	4	305	22
News & Techposts | Factorynet	4	3	1	2	2	2	2	262	45
News & Techposts | Forbes	4	4	3	1	4	2	4	310	21
News & Techposts | Gewinn	4	4	2	1	4	2	4	305	22
News & Techposts | Wirtschaftsblatt	4	4	2	1	4	2	4	305	22
Patents/Trademarks | Espacenet	1	4	4	4	4	4	3	331	8
Patents/Trademarks | USPto	1	4	3	4	4	4	3	326	10
Patents/Trademarks | EPO	1	4	3	4	4	4	3	326	10
Patents/Trademarks | TMDB	2	4	4	1	4	4	3	295	29
Scientific Publications | Science Direct	2	4	4	3	4	4	3	333	3
Scientific Publications | Scopus	2	4	4	3	4	4	3	333	3

Scientific Publications | Web of Science	2	4	4	3	4	4	3	333	3
Social Networks | Facebook	4	4	4	1	4	1	0	281	40
Social Networks | LinkedIn	4	4	4	1	4	1	0	281	40
Social Networks | Youtube	4	4	4	1	4	1	0	281	40
Social Networks | Twitter	4	4	4	1	4	1	1	286	37
Social Networks | Xing	4	4	4	1	4	1	0	281	40
Technology Reports | Gartner-TechReport	1	1	4	4	3	4	3	248	48
Technology Reports | Mc Kinsey Disruptive 12	1	1	4	4	3	4	3	248	48
Technology Reports | Boston Consulting	4	4	4	4	4	4	4	400	1
Technology Reports | Deloitte-Techtrends	1	1	4	4	3	4	3	248	48
Universities | TU-Graz	2	2	1	2	3	3	3	226	56
Universities | TU-Wien	2	2	1	2	3	3	3	226	56
Universities | MU-Leoben	2	2	1	2	3	3	3	226	56

10.3 Customized Science Direct Downloader in MatlabR2015a

```
%   Author:         Philipp Berner
%   Name:           Automated_Abstr_Test_Science Direct
%   Description:    Automated extraction of Abstracts from Publications
%                   (Sciencedirect) Ver.: 2.1
%   Date:           19.12.2016
% -------------------------------------------------------------------------
clear all
clc
options.Timeout = 30;
timeout = 30;
%-Script Variables
FileName = 'Maker-Movement';
ScanNumPage = 40;
ArtperPage = 25;
Header = sprintf('%s \r%s-Automated Summarizing of Publication-Abstracts \rDate: %s \r \r', 'Created
by Philipp Berner', FileName , date);
SLinks = fileread('Links.txt');

%-Reading extracting the links of the first 2 Pages for automated
%URL-Generation
posLinksBgn = findstr('%- %-', SLinks) +5;
posLinksEnd = findstr('-% -%', SLinks) -5;

SearchURL = cell(length(posLinksEnd), 1);

for l=1:ScanNumPage

    if l<=length(posLinksEnd)
        SearchURL{l} = SLinks(posLinksBgn(l):posLinksEnd(l));
        dataSearchPage = webread(char(SearchURL(l)));
    else
        PosArtIDBgn = findstr('name="md5" value="', dataSearchPage) +18;
        PosArtIDEnd = findstr('">', dataSearchPage(PosArtIDBgn(2):end)) -2;
        ArtID = dataSearchPage(PosArtIDBgn(2):(PosArtIDBgn(2)+PosArtIDEnd(1)));

        URL = char(SearchURL{2});
        PosChunk = findstr('chunk=', URL)+6;
        PosID = findstr('c&md5=', URL)+5;
        PosIDEnd = findstr('&', URL(PosID:end))+PosID-1;
        URLBgn = URL(1:PosChunk);
        URLMid = URL((PosChunk+1):PosID);
        URLEnd = URL(PosIDEnd:end);

        URLNew = sprintf('%s%i%s%s%s',URLBgn,l-2,URLMid,ArtID,URLEnd);
        dataSearchPage = webread(URLNew);
    end

    PosPageIDBgn = findstr('name="md5"',dataSearchPage) +18;
    PosPageIDEnd = findstr('">', dataSearchPage(PosPageIDBgn(2):end)) -2;
    nPageID = dataSearchPage(PosPageIDBgn(2):(PosPageIDBgn(2)+PosPageIDEnd(1)));

    PosTitleR = findstr('cLink artTitle S_C_artTitle', dataSearchPage);
    StringTitleR = {length(PosTitleR), 1};

    %fetch Titels
    for i = 1:length(PosTitleR)
        PosTitleBgn = findstr('">',dataSearchPage(PosTitleR(i):end)) +1;
        PosTitleEnd = findstr('</a>', dataSearchPage(PosTitleR(i):end)) -2;
        StringTitleR{i} = dataSearchPage((PosTitleR(i)+PosTitleBgn):(PosTitleR(i)+PosTitleEnd));
    end

    %<span> remove
    StringTitle = {length(StringTitleR), 1};

    for i = 1:length(StringTitleR)
        PosSpan = findstr('<span', StringTitleR{i}) -2;
        PosISpan = findstr('</span', StringTitleR{i}) -1;

        dummy = char(StringTitleR(i));
        StringTitle{i} = dummy(1:PosSpan);
             for j = 1:length(PosSpan)
                 StringTitle{i} = [char(StringTitle{i}),' ',dummy((PosSpan(j)+20):PosISpan(j))];
             end
        StringTitle{i} = [char(StringTitle{i}),dummy((PosISpan(j)+8):end)];
    end

    %fetch Abstracts
    PosAbstrURLR = findstr('extLinkBlock', dataSearchPage);
```

```
    URLAbstrR = {length(PosAbstrURLR), 1};
    StringAbstrR = {length(PosAbstrURLR), 1};

    for i = 1:length(PosAbstrURLR)
    PosAbstrURLBgn = findstr('data-url="',dataSearchPage(PosAbstrURLR(i):end)) +9;
    PosAbstrURLEnd = findstr('"', dataSearchPage((PosAbstrURLR(i)+PosAbstrURLBgn):end)) -2;
    URLAbstrR =
dataSearchPage((PosAbstrURLR(i)+PosAbstrURLBgn(1)):(PosAbstrURLR(i)+PosAbstrURLBgn(1)+PosAbstrURLEnd
(1)));
    StringAbstrR= urlread(URLAbstrR);
    PosAbstrBgn = findstr('paraText', StringAbstrR) +10;
    PosAbstrEnd = findstr('</div>', StringAbstrR(PosAbstrBgn(1):end)) -2;
    StringAbstr{i}= StringAbstrR(PosAbstrBgn:(PosAbstrBgn(1)+PosAbstrEnd(1)));
    end

    for i = 1:length(StringTitle)
        PrintArticle = sprintf('%s_Abstr%i.txt', FileName, i+(l-1)*ArtperPage);
        fid = fopen(PrintArticle, 'w');
        fprintf(fid,'%s\r\r%s', StringTitle{i},StringAbstr{i});
        fclose(fid);
    end

end
```

10.4 List of Keywords Generated by IBM Alchemy 10 Page Excerpt

Keyword List

Keyword	Abstract	Note
1904	KWfndII_AKW_knowledge-worker-occupation_Abstr354.txt	-
21st century	KWfndII_CT_cloud-computing_Abstr662.txt	-
3D computer graphics	KWfndII_3DP_additive-manufacturing_Abstr170.txt	keyword leading to a new topic
3D printing	KWfndII_3DP_3D-printing_Abstr117.txt	derived the concept of the abstract right
3T3 cells	KWfndII_AM_nanotubes_Abstr271.txt	keyword to dig deeper into a topic
454 Life Sciences	KWfndII_NGG_gene-sequencing_Abstr207.txt	-
A Brief History of Time	KWfndII_NGG_gene-sequencing_Abstr46.txt	-
A New Era	KWfndIII_AOGE_unconventional-oil_Abstr152.txt	-
Absorbed dose	KWfndIII_IT_real-time-patient-data_Abstr271.txt	-
Absorption	KWfndII_AM_nanotubes_Abstr506.txt	-
Abstraction	KWfndIII_AKW_knowledge-worker-occupation_Abstr232.txt	-
Academia	KWfndIII_MI_mobile-internet_Abstr858.txt	-
Academic journal	KWfndII_AKW_knowledge-worker-occupation_Abstr228.txt	-
Academic library	KWfndIII_MI_mobile-internet_Abstr858.txt	interesting link between topics
Academic publishing	KWfndII_AKW_knowledge-worker-occupation_Abstr228.txt	-
Amine	KWfndII_AM_graphene_Abstr406.txt	keyword to dig deeper into a topic
Amine gas treating	KWfndII_AOGE_shale-gas_Abstr209.txt	keyword to dig deeper into a topic

Ammonia	KWfndII_AOGE_shale-gas_Abstr209.txt	keyword leading to a new topic
Amplifier	KWfndIII_IT_internet-connected-sensors_Abstr106.txt	-
Anaerobic digestion	KWfndII_AOGE_shale-gas_Abstr103.txt	keyword to dig deeper into a topic
Analytic Hierarchy Process	KWfndII_3DP_direct-manufacturing_Abstr16.txt	-
Analytical chemistry	KWfndII_AM_nanotubes_Abstr211.txt	-
Analytical hierarchy	KWfndII_3DP_direct-manufacturing_Abstr16.txt	-
Anatolia	KWfndIII_RE_climate-change_Abstr972.txt	-
Anatomy	KWfndII_MI_mobile-computing-device_Abstr45.txt	-
Andropause	KWfndII_AKW_knowledge-worker-occupation_Abstr26.txt	-
Anecdotal evidence	KWfndIII_IT_real-time-patient-data_Abstr1.txt	-
Anemia	KWfndII_AM_quantum-dot_Abstr255.txt	-
Anomaly	KWfndV_IT_internet-connected-sensors_Abstr239.txt	-
Anthracite	KWfndII_AKW_knowledge-automation-tool_Abstr105.txt	-
Anthracycline	KWfndII_AM_graphene_Abstr406.txt	keyword to dig deeper into a topic
Anthropology	KWfndIV_AOGE_unconventional-oil_Abstr134.txt	-
Antioxidant	KWfndII_AM_nanotubes_Abstr211.txt	keyword leading to a new topic
AnyLogic	KWfndIII_IT_RFID-tags_Abstr250.txt	-
App Store	KWfndII_MI_mobile-internet_Abstr193.txt	-
Application programming interface	KWfndII_AR_advanced-robotic_Abstr4.txt	keyword to dig deeper into a topic
Application software	KWfndII_AKW_knowledge-automation-tool_Abstr11.txt	-
Application-specific integrated circuit	KWfndIII_IT_real-time-patient-data_Abstr261.txt	keyword leading to a new topic

Applied behavior analysis	KWfndII_AV_autonomous-guidance_Abstr430.txt	interesting link between topics
Appropriate technology	KWfndIV_ES_energy-storage-technology_Abstr359.txt	-
Aqueous solution	KWfndII_AM_graphene_Abstr430.txt	keyword to dig deeper into a topic
Aquifer	KWfndIII_AOGE_unconventional-oil_Abstr7.txt	-
Arabic language	KWfndIII_IT_internet-connected-sensors_Abstr210.txt	-
Aramid	KWfndII_3DP_additive-manufacturing_Abstr253.txt	keyword to dig deeper into a topic
Archaeology	KWfndIV_IT_internet-connected-sensors_Abstr61.txt	-
Architect	KWfndII_AV_autonomous-driving_Abstr4.txt	-
Architecture	KWfndII_AV_autonomous-driving_Abstr4.txt	-
Area	KWfndII_AR_advanced-robotic_Abstr27.txt	-
Artificial intelligence	KWfndII_AOGE_shale-gas_Abstr306.txt	keyword leading to a new topic
Artificial neural network	KWfndIII_IT_real-time-patient-data_Abstr342.txt	keyword leading to a new topic
Asia	KWfndIII_AOGE_shale-gas_Abstr208.txt	-
Assembly line	KWfndIV_IT_internet-connected-sensors_Abstr108.txt	keyword to dig deeper into a topic
Assessment	KWfndII_AR_surgical-robot_Abstr5.txt	-
Asset	KWfndIV_IT_RFID-tags_Abstr254.txt	-
Assignment problem	KWfndII_AV_autonomous-system_Abstr140.txt	-
Assumption of Mary	KWfndII_AKW_knowledge-automation-tool_Abstr925.txt	-
Astrobiology	KWfndII_NGG_synthetic-biology_Abstr17.txt	interesting link between topics
Atmosphere	KWfndIII_IT_internet-connected-sensors_Abstr240.txt	interesting link between topics
Atom	KWfndII_AM_quantum-dot_Abstr253.txt	-

Atropos scheduler	KWfndII_CT_cloud-computing_Abstr499.txt	keyword to dig deeper into a topic
Attachment theory	KWfndIII_AOGE_unconventional-reserves_Abstr18.txt	-
Attack	KWfndIII_CT_cloud-computing_Abstr882.txt	-
Attribution of recent climate change	KWfndII_RE_climate-change_Abstr468.txt	-
Augmented reality	KWfndII_3DP_direct-manufacturing_Abstr186.txt	keyword leading to a new topic
Australia	KWfndII_AKW_knowledge-worker-occupation_Abstr11.txt	-
Authentication	KWfndII_CT_cloud-technology_Abstr973.txt	keyword to dig deeper into a topic
Authorization	KWfndIII_IT_RFID-tags_Abstr149.txt	keyword to dig deeper into a topic
Auto-ID Labs	KWfndIII_IT_internet-connected-sensors_Abstr261.txt	keyword to dig deeper into a topic
Automatic identification and data capture	KWfndIII_IT_RFID-tags_Abstr310.txt	-
Automatic Identification System	KWfndIV_IT_internet-connected-sensors_Abstr108.txt	-
Automation	KWfndII_3DP_direct-manufacturing_Abstr388.txt	-
Automobile	KWfndII_AKW_knowledge-worker-occupation_Abstr207.txt	-
Automotive industry	KWfndII_3DP_direct-manufacturing_Abstr299.txt	interesting link between topics
Autonomous robot	KWfndII_AV_autonomous-driving_Abstr94.txt	interesting link between topics
Backgammon	KWfndIII_IT_RFID-tags_Abstr250.txt	interesting link between topics
Backpropagation	KWfndII_CT_cloud-computing_Abstr119.txt	-
Bacteria	KWfndII_NGG_gene-sequencing_Abstr256.txt	-
Balance sheet	KWfndIII_IT_RFID-tags_Abstr113.txt	-
Baltic Sea	KWfndII_RE_climate-change_Abstr920.txt	-
Band gap	KWfndII_AM_graphene_Abstr259.txt	keyword to dig deeper into a topic

Bankruptcy	KWfndII_3DP_direct-manufacturing_Abstr220.txt	-
Barcode	KWfndIII_IT_RFID-tags_Abstr310.txt	-
Barnett Shale	KWfndII_AOGE_shale-gas_Abstr274.txt	-
Base pair	KWfndII_AM_graphene_Abstr430.txt	-
Bathroom	KWfndV_IT_internet-connected-sensors_Abstr94.txt	-
Bathtub	KWfndV_IT_internet-connected-sensors_Abstr94.txt	-
Battery	KWfndIV_ES_electric-vehicles_Abstr353.txt	keyword to dig deeper into a topic
Battery electric vehicle	KWfndIV_ES_electric-vehicles_Abstr441.txt	keyword to dig deeper into a topic
Battery recycling	KWfndIV_ES_grid-storage_Abstr55.txt	keyword to dig deeper into a topic
Bayes' theorem	KWfndII_MI_mobile-internet_Abstr178.txt	keyword to dig deeper into a topic
Behavior	KWfndII_AV_autonomous-guidance_Abstr430.txt	-
Behaviorism	KWfndII_AV_autonomous-guidance_Abstr430.txt	-
Behavioural sciences	KWfndII_AR_advanced-robotic_Abstr61.txt	-
BET theory	KWfndII_AOGE_shale-gas_Abstr276.txt	-
Better	KWfndII_AKW_knowledge-worker-occupation_Abstr104.txt	-
Bicycle	KWfndIII_IT_internet-connected-sensors_Abstr97.txt	-
Big Five personality traits	KWfndIII_IT_RFID-tags_Abstr246.txt	-
Bill Clinton	KWfndIII_AOGE_unconventional-oil_Abstr33.txt	-
Biodiversity	KWfndIII_RE_climate-change_Abstr919.txt	-
Biofuel	KWfndIII_RE_renewable-energy_Abstr369.txt	keyword to dig deeper into a topic
Bioinformatics	KWfndII_NGG_gene-sequencing_Abstr237.txt	keyword to dig deeper into a topic

Biology	KWfndII_MI_mobile-computing-device_Abstr45.txt	interesting link between topics
Biomass	KWfndIII_RE_renewable-energy_Abstr369.txt	keyword to dig deeper into a topic
Biopsy	KWfndIII_IT_real-time-patient-data_Abstr74.txt	-
Biotechnology	KWfndII_NGG_gene-sequencing_Abstr207.txt	-
Bipolar disorder	KWfndII_AKW_knowledge-worker-occupation_Abstr26.txt	-
Bit	KWfndIII_IT_real-time-patient-data_Abstr536.txt	-
Black body	KWfndIV_ES_energy-storage-technology_Abstr719.txt	interesting link between topics
Black Sea	KWfndIII_RE_climate-change_Abstr972.txt	-
Block design	KWfndIII_IT_RFID-tags_Abstr182.txt	-
Blood	KWfndII_AM_quantum-dot_Abstr255.txt	interesting link between topics
Blood bank	KWfndIII_IT_RFID-tags_Abstr393.txt	-
Blood glucose monitoring	KWfndIII_IT_internet-connected-sensors_Abstr251.txt	interesting link between topics
Blood transfusion	KWfndIII_IT_RFID-tags_Abstr393.txt	interesting link between topics
Blood vessel	KWfndII_AM_quantum-dot_Abstr885.txt	interesting link between topics
Bluetooth	KWfndII_MI_mobile-computing-device_Abstr88.txt	-
Bone fracture	KWfndII_AOGE_shale-gas_Abstr164.txt	-
Botany	KWfndII_NGG_synthetic-biology_Abstr3.txt	-
Bragg's law	KWfndIV_IT_internet-connected-sensors_Abstr202.txt	-
Brain	KWfndIII_IT_RFID-tags_Abstr259.txt	-
Brass	KWfndII_AM_graphene_Abstr660.txt	-
Brazil	KWfndIII_AOGE_unconventional-oil_Abstr116.txt	-

Breast	KWfndIII_IT_real-time-patient-data_Abstr476.txt	-
Breast cancer	KWfndII_AM_graphene_Abstr572.txt	interesting link between topics
Breast milk	KWfndIII_IT_real-time-patient-data_Abstr360.txt	-
Breastfeeding	KWfndIII_IT_real-time-patient-data_Abstr360.txt	-
Brent Spiner	KWfndIII_IT_RFID-tags_Abstr29.txt	-
BRIC	KWfndII_3DP_additive-manufacturing_Abstr282.txt	-
British Army	KWfndII_3DP_additive-manufacturing_Abstr157.txt	-
Broadband	KWfndII_MI_mobile-internet_Abstr318.txt	-
Broadband Internet access	KWfndII_MI_mobile-internet_Abstr454.txt	-
Broker	KWfndIII_IT_real-time-patient-data_Abstr7.txt	-
Bruce Schneier	KWfndII_CT_cloud-computing_Abstr17.txt	-
Buffer solution	KWfndII_AM_quantum-dot_Abstr289.txt	-
Building	KWfndIV_ES_energy-storage-technology_Abstr725.txt	-
Building automation	KWfndV_IT_internet-connected-sensors_Abstr260.txt	-
Building Information Modeling	KWfndIII_IT_RFID-tags_Abstr273.txt	-
Bulk density	KWfndII_AOGE_shale-gas_Abstr222.txt	-
Burn	KWfndII_AKW_knowledge-worker-occupation_Abstr104.txt	-
Bus	KWfndIV_ES_energy-storage-technology_Abstr737.txt	-
Business	KWfndII_3DP_direct-manufacturing_Abstr185.txt	-
Business intelligence	KWfndIV_IT_internet-connected-sensors_Abstr61.txt	-
Business model	KWfndIII_IT_internet-connected-sensors_Abstr171.txt	-

Business models	KWfndII_MI_mobile-internet_Abstr22.txt	-
Business process	KWfndIV_IT_RFID-tags_Abstr176.txt	-
Business process modeling	KWfndIII_IT_internet-connected-sensors_Abstr171.txt	-
Business process reengineering	KWfndIII_IT_RFID-tags_Abstr354.txt	-
Business software	KWfndIII_IT_internet-connected-sensors_Abstr181.txt	-
Business terms	KWfndII_3DP_direct-manufacturing_Abstr460.txt	-
C	KWfndIII_IT_RFID-tags_Abstr95.txt	-
Cache	KWfndIII_IT_internet-connected-sensors_Abstr141.txt	-
Cadmium	KWfndII_AM_graphene_Abstr648.txt	-
Calculus	KWfndII_AOGE_shale-gas_Abstr435.txt	-
Caller ID	KWfndIII_MI_mobile-internet_Abstr568.txt	-
Canada	KWfndII_AOGE_shale-gas_Abstr350.txt	-
Cancer	KWfndII_AKW_knowledge-worker-occupation_Abstr40.txt	interesting link between topics
Cancer staging	KWfndIII_IT_real-time-patient-data_Abstr74.txt	-
Capacitor	KWfndIII_IT_internet-connected-sensors_Abstr240.txt	interesting link between topics
Capacity factor	KWfndII_3DP_direct-manufacturing_Abstr429.txt	-
Capacity utilization	KWfndII_3DP_direct-manufacturing_Abstr235.txt	-
Capital	KWfndII_AKW_knowledge-worker-occupation_Abstr13.txt	-
Capital accumulation	KWfndII_AKW_knowledge-worker-occupation_Abstr13.txt	-
Capital punishment	KWfndII_AKW_knowledge-automation-tool_Abstr971.txt	-
Car battery	KWfndIV_ES_energy-storage-technology_Abstr609.txt	keyword to dig deeper into a topic

Carbohydrate	KWfndIII_IT_internet-connected-sensors_Abstr251.txt	interesting link between topics
Carbon	KWfndII_AM_graphene_Abstr16.txt	-
Carbon capture and storage	KWfndIII_RE_climate-change_Abstr907.txt	keyword to dig deeper into a topic
Carbon dioxide	KWfndII_AM_graphene_Abstr235.txt	interesting link between topics
Carbon fiber	KWfndII_3DP_additive-manufacturing_Abstr253.txt	keyword leading to a new topic
Carbon finance	KWfndII_RE_climate-change_Abstr468.txt	-
Carbon forms	KWfndII_AM_graphene_Abstr648.txt	-
Carbon nanotube	KWfndII_AM_graphene_Abstr16.txt	keyword to dig deeper into a topic
Carcinogen	KWfndIII_AOGE_shale-gas_Abstr111.txt	-
Carcinoma in situ	KWfndIII_IT_real-time-patient-data_Abstr222.txt	-
Cardiac electrophysiology	KWfndIII_IT_real-time-patient-data_Abstr538.txt	keyword to dig deeper into a topic
Cardiology	KWfndIII_IT_real-time-patient-data_Abstr376.txt	keyword leading to a new topic
Cargo	KWfndIII_IT_RFID-tags_Abstr30.txt	-
Carnot cycle	KWfndIV_ES_energy-storage-technology_Abstr533.txt	keyword to dig deeper into a topic
Cartography	KWfndIII_AOGE_unconventional-oil_Abstr6.txt	-
Case study	KWfndII_3DP_additive-manufacturing_Abstr124.txt	-
Category theory	KWfndII_MI_mobile-computing-device_Abstr566.txt	-
Cathode	KWfndII_AM_quantum-dot_Abstr377.txt	-
Causality	KWfndII_AOGE_shale-gas_Abstr5.txt	-
Cell	KWfndII_AM_graphene_Abstr339.txt	-
Cell culture	KWfndII_AM_graphene_Abstr319.txt	-

Cell division	KWfndIII_IT_real-time-patient-data_Abstr222.txt	-
Cell nucleus	KWfndII_AM_quantum-dot_Abstr255.txt	keyword to dig deeper into a topic
Cell wall	KWfndII_AM_graphene_Abstr406.txt	keyword to dig deeper into a topic
Cellular network	KWfndII_MI_mobile-computing-device_Abstr547.txt	-
Cellular respiration	KWfndII_AM_quantum-dot_Abstr716.txt	interesting link between topics
Cellulose	KWfndII_3DP_3D-printing_Abstr45.txt	-
Central heating	KWfndV_ES_energy-storage-technology_Abstr599.txt	interesting link between topics
Central nervous system	KWfndIII_IT_RFID-tags_Abstr259.txt	-
Central processing unit	KWfndIII_IT_RFID-tags_Abstr190.txt	-
Cerebellum	KWfndIII_IT_RFID-tags_Abstr259.txt	-
Cervarix	KWfndII_AKW_knowledge-worker-occupation_Abstr40.txt	-
Cervical cancer	KWfndII_AKW_knowledge-worker-occupation_Abstr130.txt	interesting link between topics

Technology Catalog

Category	Technology	Application	Source	Publication Year	URL/Abstract
Life	-	lifelogging	Techcrunch	2016	https://techcrunch.com/2016/10/31/narrative-2/
Energy Generation	-	wave energy generator	Silicon Republic	2016	https://www.siliconrepublic.com/innovation/wave-energy-seapower-galway-bay
Energy Generation	-	water powered fuel-cell	Kickstarter	2016	https://www.kickstarter.com/projects/naturesbatterycube/the-cube-portable-water-powered-fuel-cell?ref=category_popular
Mobility	-	Hyperloop	Kickstarter	2016	https://www.kickstarter.com/projects/1629380361/waterloop-the-canadian-spacex-hyperloop-competitio?ref=category_popular
Construction	3d-printing	on site 3d-printing of houses	Interesting Engineering	2016	http://interestingengineering.com/3d-printed-office-is-the-office-of-the-future/
Advanced Production	3d-printing	carbon-fiber compounds	Science Direct	2015	Investigation into the Development of an Additive Manufacturing Technique for the Production of Fibre Composite Products
Advanced Production	3d-printing	additive manufacturing for shape memory polymer	Science Direct	2015	Characterization of polyurethane shape memory polymer processed by material extrusion additive manufacturing
Advanced Production	3d-printing	solid freeform fabrication	Science Direct	2016	The cost of additive manufacturing: machine productivity, economies of scale and technology-push

Robotics	acoustic source localization	robot orientation	Science Direct	2003	AR_service-robotics_Abstr98
Health	alginate quantum dots	gene delivery	Science Direct	2016	Cationic carbon quantum dots derived from alginate for gene delivery: One-step synthesis and cellular uptake
Computing	ambient intelligence	grid-computing	Science Direct	2014	The Internet of Things vision: Key features, applications and open issues
Mobility	antimatter propulsion	space travel	Kickstarter	2016	https://www.kickstarter.com/projects/2114765394/antimatter-propulsion?ref=category_popular
Computing	artificial intelligence	machine encryption	Techcrunch	2016	https://techcrunch.com/2016/10/28/googles-ai-creates-its-own-inhuman-encryption/
Health	artificial intelligence	intelligent health diagnostics	Silicon Republic	2016	https://www.siliconrepublic.com/start-ups/kinesis-medtech-funding
Computing	artificial intelligence	quantum cryptography	Silicon Republic	2016	https://www.siliconrepublic.com/machines/quantum-cryptography-china
Health	artificial neural network	real-time patient data	Science Direct	2016	PCV150 - Real Patients Real Data Systems
Advanced Production	astrobiology	bionics	Science Direct	2016	Industry 5.0—The Relevance and Implications of Bionics and Synthetic Biology
Life	augmented reality	google tango	Techcrunch	2016	https://techcrunch.com/2016/11/01/google-finally-launches-tango/
Life	augmented reality	Microsoft 3d	Techcrunch	2016	https://techcrunch.com/video/microsoft-gm-megan-saunders-discusses-windows-3d/5817528144c8a314ee5591dd/
Life	augmented reality	personalized books	Silicon Republic	2016	https://www.siliconrepublic.com/start-ups/cleverbooks-ar-3d-publishing-startup-week
Life	augmented reality	wearable displays	Kickstarter	2016	https://www.kickstarter.com/projects/1991375881/vufine-the-next-evolution-in-wearable-displays?ref=category_popular

Advanced Production	augmented reality	direct production	Science Direct	2016	Towards a griddable distributed manufacturing system with augmented reality interfaces
Life	augmented sound	active noise control	Kickstarter	2016	https://www.kickstarter.com/projects/1029411169/tilde-selective-noise-cancelling-earphones?ref=category_popular
Advanced Production	automation technology	longwall shearer	Science Direct	2014	Sensing for advancing mining automation automation technology development
Mobility	autonomous conductive charging	electric vehicle charging	Science Direct	2015	Implementation of autonomous distributed V2G to electric vehicle and DC charging system
Life	autonomous flying drones	security guard for infrastructure	Silicon Republic	2016	https://www.siliconrepublic.com/companies/deutsche-telekom-drone-defence-system
Energy Generation	axial-hydro-turbine	tidal hydro power	Interesting Engineering	2016	http://interestingengineering.com/meygen-worlds-largest-tidal-power-project-launched-scotland/
Health	bio-informatics	gene-sequencing	Science Direct	2016	Comparative analysis of whole genome sequencing-based telomere length measurement techniques
Health	bio-sensors	bio-markers	Science Direct	2017?? laut SD	Fluorescent biosensors enabled by graphene graphene oxide
Energy Generation	biogenic methane mining	coal-bed methane	Science Direct	2016	Biogenic methane in shale gas and coal bed methane: A review of current knowledge and gaps
Health	bioprospecting	industrializing microorganism	Silicon Republic	2016	https://www.siliconrepublic.com/innovation/extremophiles-discovery-biotech
Computing	cloak of light	photonic processors	Silicon Republic	2016	https://www.siliconrepublic.com/machines/photonics-harry-potter-invisibility-cloak
Mobility	composite cellular material	morphing airplane wings	Silicon Republic	2016	https://www.siliconrepublic.com/machines/morphing-aeroplane-wing-mit-nasa
Materials	compostable plastics	compostable shoes	Interesting Engineering	2016	http://interestingengineering.com/heeling-landfills-one-shoe-at-a-time/
Materials	condensed matter physics	quantum dot composites	Science Direct	2016	Polyaniline/carbon nanotube/CdS quantum dot composites with enhanced optical and electrical properties

Computing	dressed qubits	quantum CPU	Silicon Republic	2016	https://www.siliconrepublic.com/innovation/quantum-computer-record-10-fold-stability
Energy Generation	dye-sensitized solar cell	three-dimensional nitrogen and sulfur co-doped graphene networks	Science Direct	2016	One-step synthesis of three-dimensional nitrogen and sulfur co-doped graphene networks as low cost metal-free counter electrodes for dye-sensitized solar cells
Advanced Production	equal channel angular extrusion	3d-printing pharmaceutics	Science Direct	2016	Hot-melt extruded filaments based on pharmaceutical grade polymers for 3D printing by fused deposition modeling
Robotics	etho-robotics	service robots	Science Direct	2015	Etho robotics: What kind of behaviour can we learn from the animals?
Robotics	flexible robots	haptic technology	Science Direct	2015	A Novel Tele-Operated Flexible Robot Targeted for Minimally Invasive Robotic Surgery
Advanced Production	fused deposition molding	3d-printing electronics	Silicon Republic	2016	https://www.siliconrepublic.com/machines/3d-printed-heart-chip-harvard
Energy Generation	gale turbine	harvesting the power of typhoons and storms	Interesting Engineering	2016	http://interestingengineering.com/engineers-develop-wind-turbines-harness-typhoons/
Life	gamification	employee training	Techcrunch	2016	https://techcrunch.com/2016/11/01/axonify-raises-27m-to-gamify-employee-training-without-wasting-your-time/
Health	gene-sequencing	hearing & vision impairment	Science Direct	2016	Chapter 8 - Next Generation Sequencing in Vision and Hearing Impairment
Energy Storage	geothermal energy storage	base-load power production	Science Direct	2016	Towards the increased utilisation of geothermal energy storage
Energy Storage	gradient flow battery	smart grid	Science Direct	2016	The concentration gradient flow battery as electricity storage Technology energy dissipation
Energy Generation	Haber-Bosch process	biofuels	Science Direct	2016	A system approach in energy renewable energies sources integration in ammonia production plants
Energy Storage	heat engine	residential building with heat pump	Science Direct	2016	Cost-optimal thermal energy storage system for a residential building with heat pump heating and demand response control

Materials	hemp to graphene	super-capacitors	Techcrunch	2016	https://techcrunch.com/2016/10/27/hemp-cant-get-you-high-but-it-can-get-high-tech/
Materials	high anisotropy spin torque resonators	development of new magnetic materials	Silicon Republic	2016	https://www.siliconrepublic.com/machines/amber-transpire-research-contract
Life	holography	Microsoft Hololens	Silicon Republic	2016	https://www.siliconrepublic.com/machines/microsoft-hololens-europe
Life	holography	holographic vector display	Kickstarter	2016	https://www.kickstarter.com/projects/2029950924/holovect-holographic-vector-display?ref=category_popular
Energy Generation	horizontal drilling	shale gas exploitation	Science Direct	2017?? laut SD	Chapter Three - Exploration and Drilling in Shale Gas and Oil Reserves
Energy Generation	horizontal drilling	shale oil exploitation	Science Direct	2017?? laut SD	Chapter Three - Exploration and Drilling in Shale Gas and Oil Reserves
Energy Storage	hybrid-energy storage	super capatteries	Silicon Republic	2016	https://www.siliconrepublic.com/innovation/tyndall-national-institute-projects-cork
Mobility	hydrogen fuel-cell	e-airplane	Interesting Engineering	2016	http://interestingengineering.com/accelerating-the-future-of-aircraft-with-electricity/
Mobility	hydrogen fuel-cell	e-train	Interesting Engineering	2016	http://interestingengineering.com/worlds-first-zero-emissions-hydrogen-powered-train/
Mobility	inductive charging	electric vehicle charging	Science Direct	2015	Implementation of autonomous distributed V2G to electric vehicle and DC charging system
Energy Storage	intelligent food refrigeration	warehouses as intelligent energy hubs	Science Direct	2016	Refrigerated warehouses as intelligent hubs to integrate renewable energy in industrial food refrigeration and to enhance power grid sustainability
Life	internet connected sensors	earthquake early warning systems	Science Direct	2016	Technologies of Internet of Things applied to an Earthquake Early Warning System
Energy Storage	internet connected sensors	smart grid	Science Direct	2016	Design and implementation of a secure cloud-based billing model for smart meters as an Internet of things using homomorphic cryptography

Advanced Production	KTN-beam deflector	3d printing	Silicon Republic	2016	https://www.siliconrepublic.com/machines/3d-printing-photonics-breakthrough
Energy Generation	low-head-hydro-power	gorlov-turbine	Science Direct	2014	Theoretical and conditional monitoring of a small three-bladed vertical-axis micro hydro turbine
Computing	machine learning	digital memories with pervasive mobile devices	Science Direct	2014	Creating human digital memories with the aid of pervasive mobile devices
Health	micro electrochemical integration	wearable body sensor network	Science Direct	2016	9 - Wearable body sensor network for health care applications
Advanced Production	micro transfer printing	creation of integrated components	Silicon Republic	2016	https://www.siliconrepublic.com/innovation/tyndall-national-institute-projects-cork
Health	micro-nano-electronics	brain implanted microelectrodes	Science Direct	2014	RFID transceiver for wireless powering brain implanted microelectrodes and backscattered neural data collection
Energy Generation	microwave energy transmission	interstellar solar-energy harvesting	interesting Engineering	2016	http://interestingengineering.com/using-flying-carpets-to-light-the-world/
Materials	multi-walled carbon nanotubes	mesoporous silica	Science Direct	2016	
Nano Technology	nano crystal catalyst	H2 production	Interesting Engineering	2016	http://interestingengineering.com/splitting-water-using-tiny-nanowires/
Energy Generation	nano hybrid cathode	organic solar cells	Science Direct	2016	In situ implanting carbon nanotube-gold nanoparticles into ZnO as efficient nanohybrid cathode buffer layer for polymer solar cells
Nano Technology	nano transistors	micro CPU	Interesting Engineering	2016	http://interestingengineering.com/berkeley-makes-smallest-transistor-ever/
Nano Technology	nanorods	super conductivity	Science Direct	2016	Highly efficient yttrium-doped ZnO nanorods for quantum dot-sensitized solar cells
Computing	natural language understanding	personal shopper	Techcrunch	2016	https://techcrunch.com/2016/11/01/ibm-buys-expert-personal-shopper-from-fluid-to-build-out-watsons-conversation-skills/

Computing	natural language understanding	personal assistant	Techcrunch	2016	https://techcrunch.com/2016/11/01/rokid-the-assistant-that-can-see-hear-and-sing-raises-50m-at-450m-valuation/
Advanced Production	neutron-damage calculations	non-destructive imaging	Science Direct	2017?? laut SD	Theoretical neutron damage calculations in industrial robotic manipulators used for non-destructive imaging applications
Energy Generation	nuclear fusion	tokamak fusion reactor	Interesting Engineering	2016	http://interestingengineering.com/worlds-largest-fusion-reactor-harness-power-sun/
Nano Technology	organic modified montmorillonite (OMMT)	3d-printing nano-composites	Science Direct	2016	Mechanical and thermal properties of ABS/montmorillonite nanocomposites for fused deposition modeling 3D printing
Communication	photonics integration	fiber broadband data transmission	Silicon Republic	2016	https://www.siliconrepublic.com/innovation/tyndall-national-institute-projects-cork
Energy Generation	piezo electric materials	energy for bio implants	Silicon Republic	2016	https://www.siliconrepublic.com/innovation/tyndall-national-institute-projects-cork
Nano Technology	plasmon excited quantum dots	nanoimprinted thrombin aptasensor	Science Direct	2015	Nanoimprinted thrombin aptasensor with picomolar sensitivity based on plasmon excited quantum dots
Advanced Production	pneumatic tubular actuator	soft-grasping	Interesting Engineering	2016	http://interestingengineering.com/borrowing-natures-technology-engineer-precise-grippers/
Construction	polymer microfibers	ductile concrete	Interesting Engineering	2016	http://interestingengineering.com/new-bendable-concrete-seeks-to-be-stronger-and-durable/
Life	reusable space transporter	space tourism	Interesting Engineering	2016	http://interestingengineering.com/live-space-aboard-asgardia/
Energy Generation	reusable space transporter	interstellar helium3 mining	Silicon Republic	2016	https://www.siliconrepublic.com/innovation/long-march-5-rocket-launch-helium-3
Life	reusable space transporter	commercial space travel	Silicon Republic	2016	https://www.siliconrepublic.com/innovation/iss-expedition-49-cygnus
Energy Generation	reverse osmosis	desalination	Science Direct	2015	Integration of renewables energy system with a high share of wind and photovoltaics

Advanced Production	selective laser sintering	3d-printed magnets	Silicon Republic	2016	https://www.siliconrepublic.com/machines/3d-printed-magnets
Life	sensoric clothing	exoskeletons	Silicon Republic	2016	https://www.siliconrepublic.com/machines/ul-soft-robotics-exoskeleton
Energy Generation	spectrally selective absorber	solar vapor generator	Interesting Engineering	2016	http://interestingengineering.com/solar-vapour-generator-using-bubble-wrap/
Computing	t-ray	computer memory	Silicon Republic	2016	https://www.siliconrepublic.com/machines/t-rays-computer-memory-mipt
Health	telomere length measurement	genome sequencing	Science Direct	2016	Comparative analysis of whole genome sequencing-based telomere length measurement techniques
Advanced Production	terahertz spectroscopy	non invasive early gender definition of chickens	Techcrunch	2016	https://techcrunch.com/2016/10/30/teraegg/
Energy Generation	thermo electric materials	solar thermoelectric generator	Interesting Engineering	2016	http://interestingengineering.com/its-solar-power-but-not-as-you-know-it/
Energy Storage	vanadium redox flow battery	electric vehicle	Science Direct	2016	Assessment of the use of vanadium redox flow batteries for energy storage and fast charging of electric vehicles in gas stations
Energy Storage	vehicle to grid	smart grid	Science Direct	2016	Privacy preservation for V2G networks in smart grid: A survey
Computing	virtual machine monitoring	grid-computing	Science Direct	2016	Virtual Machine Monitoring in Cloud Computing
Life	virtual reality	gaming	Kickstarter	2016	https://www.kickstarter.com/projects/716502974/oak-turn-your-tabletop-into-an-augmented-reality-p?ref=category_popular
Advanced Production	wire-arc additive manufacturing	iron rich feal intermetallics	Science Direct	2015	Fabrication of iron-rich Fe–Al intermetallics using the wire-arc additive manufacturing process
Materials	wood-bleaching	translucent-wood	Interesting Engineering	2016	http://interestingengineering.com/scientists-can-now-create-super-strong-wooden-windows-that-dont-shatter/

10.6 Python III Script for IBM Alchemy Implementation

```
#!/usr/bin/env python3
import sys
import os
import glob
from watson_developer_cloud import AlchemyLanguageV1

def main(path):
    alchemy_language = AlchemyLanguageV1(api_key =
"4b1f080a1804c3217d7e5d2044087a882fe9dd76")

    for root, dirs, files in os.walk(path):
        for filename in files:
            filepath = os.path.join(root, filename)
            with open(filepath, 'r', encoding='utf-8', errors='ignore') as file:
                text = file.read()
                concepts = alchemy_language.concepts(text = text, language =
'english')['concepts']
                for concept in concepts:
                    print(concept['text'], end = "")
                    print(';', end = "")
                    print(filename)

if __name__ == "__main__":
    path = sys.argv[1] if (len(sys.argv) > 1) else "*"
    main(path)
```